UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN INFORMACIJSKE TEHNOLOGIJE

Magistrsko delo

Statistična analiza lojalnosti uporabnikov zavarovalniških storitev

(Statistical analysis of costumer loyalty of the insurance services)

Ime in priimek: Vida Topić
Študijski program: Matematične znanosti, 2. stopnja
Mentor: izr. prof. dr. Janez Žibert
Somentor: doc. dr. Vito Vitrih

Koper, april 2015

Ključna dokumentacijska informacija

Ime in PRIIMEK: Vida TOPIĆ

Naslov magistrskega dela: Statistična analiza lojalnosti uporabnikov zavarovalniških storitev

Kraj: Koper

Leto: 2015

Število listov: 48 Število slik: 7 Število tabel: 5

Število referenc: 26

Mentor: izr. prof. dr. Janez Žibert

Somentor: doc. dr. Vito Vitrih

UDK: 004.42(043.2)

Ključne besede: analiza podatkovne zbirke, faktorska analiza, logistična regresija, regresijsko drevo.

Math. Subj. Class. (2010): 62-07, 62N86, 62H25, 62J05

Izvleček: Osrednja tema magistrske naloge je iz podatkovne zbirke poiskati parametre, ki najbolj vplivajo na odhod zavarovancev. Za pridobitev rezultatov in enostavnejšo interpretacijo, se v prvem koraku posamično analizirajo parametri podatkovne zbirke. Testne statistike so analizirale parametre glede na odhod zavarovancev. V drugem koraku se poskuša poiskati med seboj povezane podatke, s pomočjo faktorske analize. Vsaka dobljena skupina predstavlja neko lastnost osebe. Nove skupine so uporabljene za gradnjo modelov za napovedovanje odhoda zavarovancev. V zadnjem koraku se izvede vrednotenje modeliranja. Modele vrednotimo z 10-kratnim navzkrižnim preverjanjem, nato se s testnimi statistikami analizira kakovost modelov. Omenjena analiza podatkovne zbirke in gradnje modelov je izvedena v programskem jeziku Matlab.

Key words documentation

Name and SURNAME: Vida TOPIĆ

Title of final project paper: Statistical analysis of costumer loyalty of the insurance services

Place: Koper

Year: 2015

Number of pages: 48 Number of figures: 7 Number of tables: 5

Number of references: 26

Supervisor: Assoc. Prof. Janez Žibert, PhD

Co-Supervisor: Assist. Prof. Vito Vitrih, PhD

UDC: 004.42(043.2)

Keywords: data analysis, factor analysis, logistic regression, regression tree.

Math. Subj. Class. (2010): 62-07, 62N86, 62H25, 62J05

Abstract: The main theme of this master's thesis is to search for parameters from dataset that have the most influence of customer leaving the company. In the first step, parameters from dataset are analyzed separately to acquire results that are easier to interpret. Test statistics analyze parameters based on customer leaving the company. Secondly, the factor analysis is performed to find specific groups of parameters that covariates in the same way. Each obtained group is presented with a particular property. New groups are applied for constructing models to predict if customer will leave the company. In the last step, the model evalutaion is performed by using 10-fold cross validation method, followed by test statistics to analyze the quality of models. Above mentioned dataset analysis and model construction is performed in Matlab programming language.

Kazalo vsebine

Kazalo tabel

Kazalo slik

Seznam kratic

ANOVA	analiza variance
AUC	ploščina pod ROC krivuljo
CART	kategorijska in regresijska drevesa
LSD	test najmanjše značilne razlike
MSR	povprečna vsota kvadratov napake
PCA	metoda glavnih komponent
ROC	karakteristika delovanja sprejemnika

Zahvala

Posebna zahvala gre mentorju izr. prof. dr. Janezu Žibertu, ki je v meni vzbudil zanimanje za podatkovno rudarjenje in me v času pisanja magistrskega dela usmerjal in potrpežljivo popravljal.

Zahvalila bi se somentorju doc. dr. Vitu Vitrihu za izčrpne popravke in komentarje, ki so mi veliko pripomogli pri pisanju magistrskega dela.

Prav posebna zahvala gre tudi mentorju na strani zavarovalnice mag. Janezu Kralju, ki mi je pomagal pri interpretaciji rezultatov.

Mož Duško, mama, oče, brat in najpomembnejši novi član družine sin Nikola, so osebe, katerim bi se zahvalila za podporo na vsakem življenjskem koraku.

Prijateljici Marija in Nastja, hvala vama za nepozabna študijska leta.

1 UVOD

Magistrsko delo sodi na področje podatkovnega rudarjenja. Podatkovno rudarjenje vključuje računske postopke, s katerimi odkrivamo vzorce v podatkih. Cilj procesa je pridobiti informacije iz podatkovne zbirke in jih pretvoriti v razumljive strukture za nadaljnjo uporabo.

Z rastjo in razvojem gospodarstva se je število ponudnikov storitev večalo, medtem ko količina povpraševalcev stagnira in s tem raste konkurenca med podjetji. Konkurenca je neizogibna in v večini primerov zaželjena. Konkurenca na trgu pomeni možnost vstopa novih ponudnikov na trg, racionalnost tržnih osebkov, optimizacijo proizvodnih procesov kar vse se odraža na ceni proizvoda. Posledica tega pojava je odhajanje obstoječih in težavno pridobivanje novih strank. V ta namen so podjetja poleg splošnega oglaševanja, ki nagovarja celotno populacijo, začela izvajati posebne ponudbe in akcije, ki so bile namenjene le izbranemu naboru strank. Začetki pametnih izbir potencialnih strank so se začeli s pošiljanjem ponudbe izbranim osebam, ki so bile določene po nekem kriteriju (starost, spol, lokacija bivanja itd.). Ker so podjetja s takim načinom trženja pridobivala večji odziv na ponudbo kot pri naključno poslanih ponudbah, so začela izpopolnjevati model. Prvi pogoj za izpopolnjevanje izbire potencialnih strank je pridobivanje podrobnejših informacij o strankah. V ta namen so postopoma uvedli zbiranje osebnih podatkov, beleženje nakupov obstoječe stranke in ostale navade, ki jih je moč pridobiti iz nakupa. Vsaka od strank je imela svoj način obnašanja, vendar se je hitro opazilo, da je stranke možno razvrščati v več skupin. V vsako skupino so porazdeljeni posamezniki, ki imajo podoben način obnašanja in se podobno odzovejo ob posebnih priložnostih. Pripadnike take skupine dojemamo kot en tip stranke, ki ji na podlagi njenih preferenc in zmožnosti ponudimo posebej prilagojeno ponudbo. Ker obnašanje strank postaja vedno bolj kompleksno, saj je podatkov vedno več, je analiza le-teh postala bolj zahtevna in posledično ne več intuitivna. Posledično je potrebno za pravilno razčlenitev trga strank in za pridobivanje znanja iz zbranih podatkov uporabljati matematične metode podatkovnega rudarjenja.

Splošno znano je, da podjetje enostavneje obdrži obstoječo stranko, kot pridobi novo. Kazalnika, ki kažeta na to, sta nasičenost trga in ekonomska kriza, ki zmanjšata nabor potencialnih strank. Zato so se podjetja odločila razvijati modele, ki napovedujejo odhod stranke v konkurenčno podjetje.

V magistrskem delu smo se osredotočili na panogo zavarovalništva, v kateri so z zakoniki kot so: Zakon o varstvu konkurence, Zakon o preprečevanju omejevanja konkurence, Zakon o varstvu potrošnikov, Zakon o zavarovalništvu in Obligacijski zakonik, zastavljena pravila, katerim mora zadostiti vsaka zavarovalnica. V zakonikih so zavedene konkretne prepovedi omejevanja konkurence, kot so: dogovarjanja o višini premije, o pogojih za opravljanje zavarovalnih storitev, o medsebojnem nadziranju in podobno. Zaradi zakonskih omejitev in temeljnih matematičnih pravil, s katerimi je izračunana minimalna višina premije, s katero zavarovalnica krije potencialne škode, je iz zavarovalniškega stališča ključnega pomena zadovoljstvo stranke. Hipoteza magistrskega dela je:"Ali znamo napovedati karakteristike zavarovanca, ki bo opustil zavarovanje?" Podatke smo pridobivali iz obstoječih zbirk podatkov o zavarovancih.
V prvem poglavju smo opisali strukturo podatkovne zbirke, razložili proces pridobivanja značilk in postopek analize podatkovne zbirke. Opisan je tudi proces modeliranja podatkov in potek vrednotenja zgrajenih modelov. Drugo poglavje je namenjeno predstavitvi dobljenih rezultatov. V zadnjem, tretjem poglavju smo povzeli rezultate in jim pripisali pomen.

1.1 Pregled področja

V članku [19] je zajeta večina raziskav med letom 2000 in 2006, katere so uporabljale tehnike podatkovnega rudarjenja na področju upravljanja odnosov s strankami (CRM). CRM delimo na štiri skupine:

- Opredelitev stranke.
- Zanimivost stranke.
- Zadržanje stranke.
- Razvoj stranke.

S temi skupinami lahko predstavimo sistem za upravljanje odnosa s strankami. Cilj vsake skupine je zbiranje podatkov in s tem pripomoči k boljšemu razumevanju stranke. Tehnike podatkovnega rudarjenja lahko pomagajo pri doseganju tega cilja tako, da poiščejo prostemu očesu skrite vzorce strankinega obnašanja in strankinih lastnosti iz podatkovne baze. Podatkovno rudarjenje iz podatkov zgradi model. Vsaka tehnika podatkovnega rudarjenja je lahko vsebovana v enem ali več naštetih tipov podatkovnega modeliranja:

- Združevanje (ang. *Association*) uporabimo, kadar želimo odkriti, med katerimi parametri znotraj zapisa obstaja povezava. Običajno se tak tip podatkovnega

rudarjenja uporablja pri tržni analizi nakupovalne košarice in navzkrižni prodaji produktov. Orodja, s katerimi izvajamo ta tip modeliranja, sta testne statistike in apriori algoritmi.

- Klasifikacija (ang. *Classification*) je ena od najpogostejših učnih modelov v podatkovnem rudarjenju. Za razvrščanje se največkrat uporabi odločitveno drevo.

- Rojenje [5] (ang. *Clustering*) poskuša heterogeno populacijo razvrstiti v več homogenih rojev. Od klasifikacije se razlikuje v tem, da je število in tip rojev neznano.

- Napovedovanje (ang. *Forecasting*) določi pričakovano vrednost v naslednji časovni enoti glede na pretekle vzorce vedenja. Tehniki podatkovnega rudarjenja za napovedovanje sta nevronske mreže in analiza preživetja.

- Regresija (ang. *Regression*) je ena izmed tehnik statističnega ocenjevanja. Uporablja se za preslikavo vsakega podatkovnega objekta v realno vrednost. Tehniki, ki ju najpogosteje uporabljamo, sta linearna regresija in logistična regresija ([12], [21]).

- Odkrivanje zaporedja (ang. *Sequence discovery*) je zaznavanje povezanih vzorcev skozi neko časovno obdobje. Cilj tega modela je napovedati trende in določiti faze procesa ter zaznati nenavadne vzorce.

- Vizualizacija (ang. *Visualization*) je predstavitev podatkovne zbirke uporabnikom, tako da je lahko videti vzorce v podatkih.

Še veliko tehnik podatkovnega rudarjenja je preučenih:

- Metoda podpornih vektorjev in model nevronskih mrež sta tehniki podatkovnega rudarjenja, ki sta bili uporabljeni v članku [3]. Avtor ju uporabi za odkrivanje potencialnih odhodov strank. Izkaže se, da je metoda podpornih vektorjev boljša, saj je bolj učinkovita pri razvrščanju in manj občutljiva na šum v podatkih.

- Metoda naključnih gozdov je uporabljena v članku [25]. V članku so uporabili izboljšano metodo učenja, ki se imenuje izboljšana uravnotežena metoda naključnih gozdov (IBRF) in predstavili njegovo uporabo na dejanskih podatkih. Izkaže se, da je uporabljena izboljšana metoda klasifikacije strank boljša od nevronskih mrež, odločitvenih dreves in metode podpornih vektorjev.

V članku [21] avtorji odgovorijo na vprašanja, kot so: kako ugotoviti katere spremenljivke prinesejo modelu največ informacije, relativni pomen spremenljivke in kako implementirati model v posel. V članku je predstavljen razvoj modela z uporabo logistične

regresije. Model išče ponavljajoče vzorce, ki nakazujejo na strankino nezadovoljstvo in možen odhod. Model je bil zgrajen na podlagi zadnjega stanja stranke in cilj je bilo napovedati status stranke, označen kot aktiven oziroma preklican. Uporabljene niso bile samo police, ki so prenehale veljati v raziskovanem letu, ampak tudi pretekla leta, zato da zadosti posplošitvi podatkov na časovno obdobje in zajame demografske spremembe. Da bi model ustvarjal manjšo napako napovedovanja odhoda stranke, so nad podatki v prvem koraku izvršili rojenje ter zavarovance razdelili v dva roja. V naslednjem koraku so nad vsakim rojem aplicirali logistično regresijo, katera je določila verjetnost odhoda. Rezultati članka so pokazali, da je za stalnost stranke ključnega pomena, koliko časa je stranka v dotični zavarovalnici, koliko časa je pri zadnjem produktu ter koliko časa traja veljavna polica.

Članek [9] predstavi teoretične osnove modeliranja z logistično regresijo na primeru zavarovalniških podatkov. V prvem koraku mora raziskovalec izbrati atribute, ki dajejo pomen napovedani vrednosti. Očitno je, da bo zavarovanec z 20 letno zgodovino na zavarovalnici bolj zvest kot zavarovanec, ki je šele prišel. Torej v tem primeru leta pomenijo veliko pri napovedovanju odhoda zavarovanca. V drugem koraku je potrebno oceniti interakcijo med atributoma. Včasih se lahko zgodi, da kombinacija vrednosti dveh atributov, ki delujeta skupaj, sproži drugačen odziv, kot če bi vsak od teh atributov deloval posebej. Potrebno je tudi paziti na neodvisnost med atributi in izločiti atipične zapise, ki močno odstopajo od množice, saj lahko le-ti močno popačijo model. V raziskovalnem delu je avtor prišel do ugotovitve, da so stranke, ki imajo zavarovano motorno vozilo, zelo dovzetne za odhod. Na drugi strani so zavarovanci, ki imajo vsaj dve aktivni polici premoženjskega zavarovanja, ki se izkažejo za zelo zveste zavarovance, saj je verjetnost, da bodo odšli, zelo majhna.

2 METODOLOGIJA DELA

V magistrskem delu smo poskusili analizirati podatkovno zbirko zavarovancev, ki je sestavljena iz oseb, ki so zavarovanje prekinili, in oseb, ki imajo še vedno aktivno zavarovalno polico. Cilj analize je bil zgraditi matematični model, ki bo znal prepoznati zavarovanca, ki bo zavarovanje prekinil, in opredeliti njegove karakteristike.

2.1 Predstavitev podatkovne zbirke

Pred analizo podatkov je potrebno pripraviti podatke, ki najbolje opišejo problem, s katerim se spopadamo. Bistvo tega koraka je v model vpeljati tiste parametre, ki bodo modelu prinesle kar se da več informacije. Takim parametrom pravimo v jeziku podatkovnega rudarjenja značilke ali atributi. Le tako bomo dobili kvalitetne in razumljive rezultate.
V našem primeru smo zgradili podatkovno zbirko iz sedemnajstih parametrov. Vsak zapis predstavlja enega zavarovanca. En zapis poseduje sedemnajst lastnosti zavarovanca. Podatkovna zbirka zajema vzorce zavarovancev, ki so bodisi imeli, bodisi imajo aktivno avtomobilsko zavarovanje. Naključno smo zajeli 44967 zavarovancev pod sledečimi pogoji:

- zavarovanec je imel sklenjeno vsaj eno avtomobilsko zavarovanje,
- zavarovanec je fizična oseba,
- zavarovanec mora biti star vsaj 18 let,
- zavarovanec mora biti v enem izmed statusov: brezposeln, študent, dijak, upokojenec, zaposlen. Izločili smo osebe s statusom umrl.

Prvih osem parametrov, ki so vključeni v podatkovno zbirko, smo zgradili tako, da smo zajeli trenutni status osebe ter šteli dogodke iz zavarovančeve zgodovine. Prva dva parametra, katera smo vključili v podatkovno zbirko, sta spol in starost. V podatkovno zbirko je vključen tudi parameter imenovan status, ki označuje trenutno aktivnost osebe. Možni statusi so: brezposeln, študent, dijak, upokojenec, zaposlen. Četrti in peti parameter štejeta, koliko polic je zavarovanec sklenil. Naslednji parameter bo štel, preko koliko prodajnih poti je zavarovanec sklepal avtomobilska zavarovanja. Z besedo

Tabela 1: Pregled zbranih parametrov.

Kategorijski parametri	spol,
	status
Zvezni parametri	starost,
	število avtomobilskih polic,
	število vseh polic,
	število prodajnih poti,
	število zavarovalniških stebrov,
	število let pri zavarovalnici,
	število let pri avtomobilskem zavarovanju,
	število škodnih zahtevkov,
	vsota vseh likvidiranih škod,
	vsota likvidiranih škod avtomobilskega zavarovanja,
	vsota vseh vplačanih premij,
	vsota vplačanih premij avtomobilskega zavarovanja,
	število let z vsaj enimi škodnim dogodkom,
	število škodnih dogodkov v prvih dveh tretjinah in
	število škodnih dogodkov v zadnji tretjini
	časovne premice.

prodajna pot označujemo način, preko katerega je zavarovanec sklenil zavarovanje. Možne prodajne poti so: agencije, poslovna enota, elektronska prodaja (sklepanje zavarovanj preko spleta), mobilna prodaja (sklepanje zavarovanj prek mobilnih naprav), itd. Parameter število prodajnih poti bo nakazoval na fleksibilnost zavarovanca, oziroma bo v nasprotnem primeru izkazal nezadovoljstvo nad raznimi načini sklepanja. Sedmi parameter šteje, v koliko zavarovalniških stebrih je zavarovanec vključen. Zavarovalniške stebre smo definirali tako, da smo zavarovanja grupirali v štiri skupine. Grupirali smo avtomobilska zavarovanja, življenjska zavarovanja, zdravstvena zavarovanja in premoženjska zavarovanja. Osmi in deveti parameter, ki smo ju vključili v podatkovno zbirko, štejeta, koliko let je zavarovanec na zavarovalnici oziroma koliko let je bil aktiven na področju avtomobilskega zavarovanja. Deseti parameter šteje škodne zahtevke. S tem parametrom poskušamo opisati frekvenco uveljavljanja škodnih dogodkov pri posameznem zavarovancu. Naslednji štirje parametri seštevajo vplačana in izplačana sredstva. Prva dva sta vsota likvidiranih škod iz naslova avtomobilskega zavarovanja ter vsota vseh likvidiranih škod. Druga dva parametra opisujeta količino vplačanih sredstev v avtomobilsko zavarovanje in vsa ostala zavarovanja. Petnajsti

parameter, ki smo ga vključili v podatkovno zbirko, označuje število let, ki so imela vsaj en škodni dogodek. Šestnajsti in sedemnajsti parameter opisujeta dogajanje na določenem časovnem intervalu zgodovine zavarovanca. Šestnajsti parameter prešteje število škodnih dogodkov v prvih dveh tretjinah časovne premice. Šestnajsti parameter pa šteje škodne dogodke, ki so se zgodili v zadnji tretjini.

2.2 Izpeljava in analiza značilk

Podatkovna zbirka, katero smo analizirali, sestavlja štirinajst numeričnih in dve kategorijski značilki. V nadaljevanju smo z izpeljavo opisnih značilk določili gradnike, s katerimi smo gradili model.

2.2.1 Izpeljava značilk

Naslednji korak pri gradnji podatkovne zbirke je pridobivanje značilk z zbranimi parametri. Značilke, kot so spol, starost, status, število let pri avtomobilskem zavarovanju, število let pri zavarovalnici, število prodajnih poti in število zavarovalniških stebrov, ne potrebujejo nikakršne obdelave in jih dobimo neposredno iz zbranih parametrov. Označujemo jih z oznakami spol, starost, status, št. let avto, št. let, št. poti in št. stebrov.
Da bi zadostili pogoju neodvisnosti med značilkami, smo naslednje značilke normirali. Značilka z oznako št. škod bo izražala povprečno frekvenco uveljavljanja škodnih zahtevkov na letni ravni na eno avtomobilsko polico. Značilko smo pridobili tako, da smo število škodnih zahtevkov iz naslova avtomobilskih zavarovanj delili s produktom števila let na avtomobilskem zavarovanju in številom sklenjenih avtomobilskih polic:

$$\text{št. škod} = \frac{\text{št. škodnih zahtevkov}}{(\text{št. let avto. zavarovanja} * \text{št. avtomobilskih polic})}. \tag{2.1}$$

Naslednje štiri značilke predstavljajo povprečno vsoto izplačil oziroma vplačil na polico. Prvi dve značilki označujeta povprečna vplačila in izplačila, vezana na avtomobilsko zavarovanje. Značilka označena kot avto. premija, označuje povprečno vplačano premijo na polico v obdobju enega koledarskega leta. Pridobljena je iz razmerja med parametrom, ki šteje vsoto vplačanih premij, in produktom parametra, ki šteje leta avtomobilskih zavarovanj, s številom sklenjenih veljavnih avtomobilskih polic.

$$\text{avto. premija} = \frac{\text{vsota vplačanih premij}}{(\text{št. let avto. zavarovanja} * \text{št. avtomobilskih polic})}. \tag{2.2}$$

Preostale tri značilke so pridobljene na podoben način kot zgoraj opisana značilka avto. premija. Značilka, označena kot avto. škode, označuje povprečno izplačano škodo

v obdobju enega leta. Drugi dve značilki opisujeta povprečna vplačila in izplačila, vezana na vse veljavne sklenjene police. Značilko, ki opisuje povprečno izplačano škodo na polico v obdobju enega koledarskega leta, označujemo z oznako škode. Značilko, ki opisuje povprečno vplačano premijo na polico v obdobju enega koledarskega leta, označujemo z oznako premija.
Naslednja značilka izraža aktivnost zavarovanca tako, da prešteje povprečno število sklenjenih polic na leto. Značilko označimo z oznako št. polic.

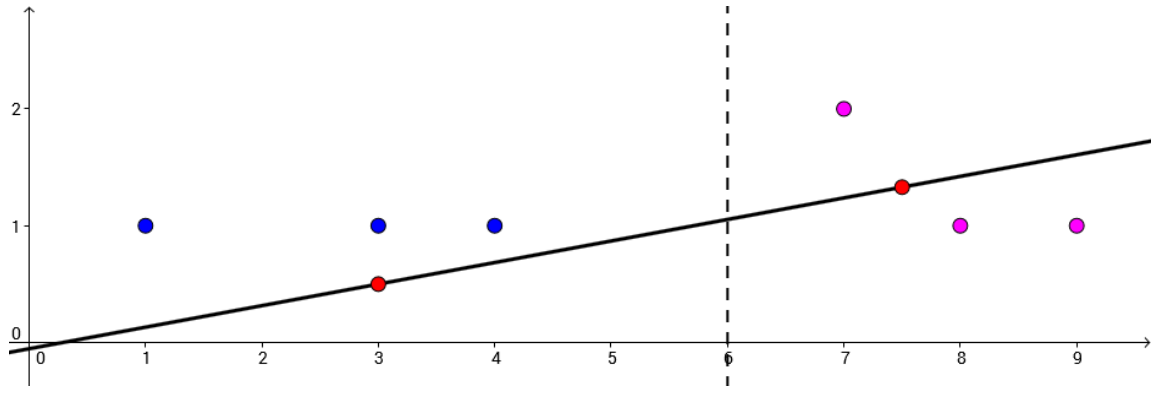
$$\text{št. polic} = \frac{\text{št. sklenjenih polic}}{\text{št. let na zavarovalnici}}. \tag{2.3}$$

Značilka, ki odraža delež škodnih let v zgodovini avtomobilskega zavarovanja zavarovanca, je dobljena iz parametrov število let avtomobilskega zavarovanja in število škodnih let avtomobilskega zavarovanja. Značilko smo označili kot leta škod.

$$\text{leta škod} = \frac{\text{št. let avto. zavarovanja}}{\text{št. let z vsaj eno prijavljeno škodo}}. \tag{2.4}$$

Zadnja, šestnajsta značilka, imenovana naklon, bo opisovala trend naraščanja oziroma padanja škodnih dogodkov skozi čas. Taki značilki pravimo dinamična značilka [6]. Dinamična značilka je odvisna od vrednosti drugih parametrov v nekem časovnem intervalu. V našem primeru je to razmerje med številom škodnih dogodkov v prvih dveh tretjinah zgodovine zavarovanca in številom škodnih dogodkov v zadnji tretjini njegove zgodovine na zavarovalnici.

$$\text{naklon} = \frac{\text{št. škod v zadnji tretjini}}{\text{št. let v zadnji tretjini}} : \frac{\text{št. škod v prvih dveh tretjinah}}{\text{št. let v prvih dveh tretjinah}}. \tag{2.5}$$



Slika 1: Modre točke označujejo število škod v prvih dveh tretjinah. Roza točke označujejo število škod v zadnji tretjini. Rdeči točki pa povprečje škod na leto. Premica nakaže na trend rasti.

Zaradi lažjega razumevanja si poglejmo en primer. Zavarovanec je pri zavarovalnici polnih 9 let. Kar moramo narediti, je prešteti, koliko škodnih dogodkov je uveljavljal v

prvih šestih letih ter prešteto število normirati tako, da dobimo povprečno število škod na leto. V naslednjem koraku moramo prešteti koliko škodnih dogodkov v povprečju je imel v zadnjih treh letih ter vsoto normirati tako, da dobimo povprečje škod na leto. V kolikor je količnik (2.5) večji od 1, je trend uveljavljanja škod rasel, v kolikor pa je manjši od 1, je trend škod padal.
V našem primeru imamo eno odzivno kategorijsko značilko, ki nakaže, ali je zavarovanec prekinil zavarovalniško razmerje, ali ne. Na podlagi te smo gradili naše modele.

2.2.2 Analiza značilk

Analizo značilk smo izvedli z namenom odkrivanja dodatnih informacij o značilkah, ki bodo pripomogle pri gradnji modela. Značilke, ki smo jih vključili v podatkovno zbirko, smo gradili z namenom, da najbolje opišemo problem odhoda zavarovancev. Z analizo smo dobili dodatno informacijo o moči selekcije značilke, ki deli zavarovance na aktivne in neaktivne.
Analiza podatkovne zbirke prispeva veliko k razumevanju in oceni kvalitete zbranih podatkov. Ocenili smo, ali obstaja signifikantna razlika med osebami, ki pripadajo različnim razredom odzivne značilke. Analizirali smo podatke glede na to, ali je zavarovanec prekinil zavarovalno razmerje ali ne. Uporabili smo metodologijo preverjanja hipotez, kjer smo za vsako značilko preverjali hipotezo, ali obstajajo razlike med vzorci zavarovancev, ki so prekinili razmerje in tistimi, ki niso prekinili. Hipoteze smo testirali po standardnih postopkih z uporabo ustreznih statističnih testov.
Lilliefors test [16] je dvostranski test, ki preverja, kako dobro se vzorec prilagaja družini normalnih porazdelitev. V članku [16] je predstavljena tabela in osnovni princip delovanja Lilliefors testa. Statistika testa je definirana s konstanto

$$D = \max_X | F^*(X) - S_N(X) |,$$

kjer je $S_N(X)$ funkcija kumulativne porazdelitve množice meritev X in $F^*(X)$ je funkcija kumulativne normalne porazdelitve vzorca s povprečjem vzorca in varianco vzorca. Če vrednost D preseže kritično vrednost tabele, se ničta hipoteza, da množica meritev X izhaja iz normalne porazdelitve, zavrne. Iz česar sledi, da porazdelitev vzorca ne pripada družini normalne porazdelitve.
Wilcoxon-ov test vsote rangov [18] je neparametrični test, ki ga lahko uporabimo, kadar imamo dva vzorca z ordinalnimi vrednostmi, ki nista normalno porazdeljena, sta neodvisna in želimo ugotoviti ali se mediani rangov vzorcev statistično signifikantno razlikujeta. Wilcoxonov test ni občutljiv na odstopajoče točke (ang. *outlier*), saj meritve vseh vzorcev razvrsti v eno vrsto po velikosti naraščajoče in jim dodeli rang. Najmanjša meritev dobi rang 1. V kolikor sta v vrsti dve ali več enakih meritev, potem vsaki od meritev pripišem povprečni rang enakih meritev.

Izbira testne statistike je odvisna tudi od tipa spremenljivke, ki jo testiramo. V primeru, ko testiramo kategorijske lastnosti populacije, se lahko poslužujemo χ^2 testa, ki je natančneje opisan v knjigi [18]. χ^2 test odgovori na vprašanje, ali je pričakovana porazdelitev spremenljivke enaka izmerjeni porazdelitvi spremenljivke.
Statistika, s katero izmerimo, ali se izmerjena in pričakovana porazdelitev razlikuje je:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{2.6}$$

kjer je O_i izmerjena porazdelitev in E_i pričakovana porazdelitev spremenljivke v razredu i.
Z uporabo Lilliefors-ovega testa se je izkazalo, da značilke niso normalno porazdeljene, zato smo uporabili neparametrični test, tako imenovani Wilcoxon-ov test. Ničta hipoteza testa: Mediani rankov vzorcev sta enaki. Alternativna hipoteza testa: Mediani rankov vzorcev se razlikujeta.
V primeru kategorijskih podatkov smo uporabljali χ^2-test, ki primerja razmerje med deleži. Ničta hipoteza testa: Pričakovana porazdelitev se ne razlikuje od izmerjene porazdelitve vzorca.
Ničto hipotezo smo zavrnili, če je p-vrednost izbrane testne statistike manjša od 0.05. P-vrednost je verjetnost, da testna statistika ob predpostavki, da ničta hipoteza velja, zavzame vrednost, ki je večja ali enaka izračunani testni statistiki iz vzorca. P-vrednost pove več kot statistična značilnost. Pri statistični značilnosti povemo, da smo npr. v primeru napake zavrnitve 5% sprejeli hipotezo. Tu pa še izvemo, kolikšna je verjetnost, da bi ob izračunani testni statistiki naredili napako zavrnitve hipoteze.

2.3 Predstavitev in gradnja modelov

V nadaljevanju smo poskušali poiskati skupine podatkov, ki so med seboj povezane. To smo izvedli s postopkom faktorske analize. Predstavitev podatkovne zbirke s faktorji smo uporabili za gradnjo modelov za napovedovanje prekinitve zavarovalniških razmerij.

2.3.1 Faktorska analiza

Faktorsko analizo [14] smo v našem primeru izvedli z uporabo analize glavnih komponent (PCA) [22] z ustrezno rotacijo.
Skupna točka metode glavnih komponent in faktorske analize je reduciranje podatkov in dimenzije prostora. Cilj faktorske analize je opisati lastnosti, ki niso merljive (npr. družbeni status) in povzročajo povezanost med spremenljivkami, ki smo jih zajeli v

podatkovno zbirko z neposrednimi meritvami (npr. število avtomobilov, mesečni prihodek, itd.). Na tak način model poskuša pojasniti povezave med večjim številom spremenljivk, ki podobno kovarirajo, z manjšim številom faktorjev. Metoda PCA znotraj podatkovne zbirke poišče smeri največje variance in jih uporabi za razvrstitev podatkov v nekaj dimenzij. Tako poišče množice spremenljivk, ki so si med seboj pomensko povezane.

Z metodo faktorske analize želimo ugotoviti, ali se zveze med spremenljivkami, ki smo jih pridobili neposredno z meritvami, lahko pojasni z manjšim številom posredno opazovanih faktorjev.

V našem primeru smo poskušali poiskali povezavo med 14-timi numeričnimi značilkami, ter predstaviti isti prostor z manjšim številom faktorjev in najmanjšo možno rekonstrukcijsko napako.

Ideja PCA metode je preslikati višji n dimenzionalni prostor v nižji m dimenzionalni prostor, tako da bo rekonstrukcijska napaka minimalna. Metoda tvori nov prostor spremenljivk, ki ga definirajo glavne komponente. Glavne komponente so med seboj ortogonalne, tvorijo bazo m dimenzionalnega prostora in vsako meritev iz začetnega prostora spremenljivk lahko zapišemo z linearno kombinacijo glavnih komponent. Geometrijsko gledano so glavne komponente koordinatne osi prostora $\mathbb{R}^m$. Algebrsko, pa lahko glavne komponente predstavimo z linearno kombinacijo slučajnih spremenljivk $x_i : x_1, x_2, \ldots, x_n$, $X = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$. Definirajmo linearno preslikavo $L : \mathbb{R}^n \to \mathbb{R}^m$, za katero velja

$$X \mapsto F^T X + b =: Z = (z_1, z_2, \ldots, z_m)^T \in \mathbb{R}^m, \tag{2.7}$$

kjer je F ortonormirana projekcijska matrika iz $\mathbb{R}^{n\times m}$ in b korekcijski vektor iz prostora $\mathbb{R}^m$.

V kolikor predhodno vsaki spremenljivki x_i odštejemo pričakovano vrednost μ_i, oziroma centriramo prostor $\mathbb{R}^n$, lahko linearno preslikavo L predstavimo tudi kot

$$X \mapsto F^T X =: Z. \tag{2.8}$$

Ko poskušamo iz vektorja Z nazaj pridobiti centriran vektor X, se poslužujemo povratne projekcije $L^{-1} : \mathbb{R}^m \to \mathbb{R}^n$

$$Z \mapsto FZ = FF^T X =: \tilde{X}. \tag{2.9}$$

Vrednosti vektorja $\tilde{X}$ so zaradi preslikave F izpostavljene rekonstrukcijski napaki $r := r(X, \tilde{X})$. Rekonstrukcijska napaka r je definirana kot vsota razlik spremenljivk x_i in preslikanih spremenljivk $\tilde{x_i}$ s povratno projekcijo:

$$r(X, \tilde{X}) := \sum_{i=1}^{n} E[(x_i - \tilde{x_i})^2]. \tag{2.10}$$

Cilj metode je poiskati tako preslikavo, da bo rekonstrukcijska napaka najmanjša.

Lema 2.1. *Velja*

$$\sum_{i=1}^{n} E[x_i^2] = \sum_{i=1}^{n} E[\tilde{x_i}^2] + \sum_{i=1}^{n} E[(x_i - \tilde{x_i})^2]. \tag{2.11}$$

Dokaz. Za vsako meritev $x = X(\omega)$, zaradi lastnosti linearne preslikave, velja:

$$\|x\|^2 = \|\tilde{x}\|^2 + \|x - \tilde{x}\|^2 \quad \text{oz.} \quad \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} \tilde{x_i}^2 + \sum_{i=1}^{n} (x_i - \tilde{x_i})^2. \tag{2.12}$$

Iz enakosti (2.12) z uporabo povprečne vrednosti sledi:

$$\sum_{i=1}^{n} E[x_i^2] = \sum_{i=1}^{n} E[\tilde{x_i}^2] + \sum_{i=1}^{n} E[(x_i - \tilde{x_i})^2]. \tag{2.13}$$

□

Lema 2.2. *Velja*

$$\sum_{i=1}^{n} E[\tilde{x_i}^2] = \sum_{i=1}^{m} E[z_i^2]. \tag{2.14}$$

Dokaz. Ker za $\forall \tilde{x} = FF^T X(\omega)$ velja:

$$\tilde{x} = \sum_{i=1}^{m} f_i z_i, \tag{2.15}$$

sledi

$$\sum_{i=1}^{n} \tilde{x_i}^2 = \sum_{i=1}^{n} \|\tilde{x}\|^2 = \sum_{i=1}^{m} \|f_i z_i\|^2 = \sum_{i=1}^{m} \|f_i\|^2 z_i^2 = \sum_{i=1}^{m} z_i^2.$$

Pri tem smo uporabili dejstvo, da je F ortonormirana baza. Uporabimo pričakovane vrednosti in vidimo, da velja:

$$E[\sum_{i=1}^{n} \tilde{x_i}^2] = \sum_{i=1}^{n} E[\tilde{x_i}^2] = \sum_{i=1}^{m} E[z_i^2].$$

□

Če povzamemo do sedaj izpeljana dejstva vemo, da je

$$E[z_i] = E[f_i^T X] = f_i^T \sum_{i=1}^{n} E[x_i] = f_i^T \mu,$$

kjer z $\mu := (\mu_1, \mu_2, \ldots, \mu_n)^T$ označujemo vektor povprečij spremenljivk x_i. Vemo tudi, da je

$$\mathrm{Var}(z_i) = E[z_i^2] - E^2[z_i].$$

Torej velja:

$$E[z_i^2] = E^2[z_i] + \mathrm{Var}(z_i)$$
$$\sum_{i=1}^{n} E[\tilde{x_i}^2] = \sum_{i=1}^{m} E[z_i^2] = (f_i^T \mu)^2 + \sum_{i=1}^{m} \mathrm{Var}(z_i).$$

Da bi lažje interpretirali minimizacijo rekonstrukcijske napake r predpostavimo, da smo začetni prostor $\mathbb{R}^n$ centrirali in je zaradi tega povprečje $\mu^T = (0, \cdots, 0)$. Iz tega sledi:

$$\sum_{i=1}^{n} E[\tilde{x_i}^2] = \sum_{i=1}^{m} \mathrm{Var}(z_i). \tag{2.16}$$

Z uporabo enačbe (2.11) in enačbe (2.16) lahko sklepamo sledeče:

$$r = \sum_{i=1}^{n} E[(x_i - \tilde{x_i})^2] = \sum_{i=1}^{n} E[x_i^2] - \sum_{i=1}^{m} \mathrm{Var}(z_i). \tag{2.17}$$

Ker je $E[x_i^2]$ določen in želimo minimizirati $E[(x_i - \tilde{x_i})^2]$ je to enako, kot da bi maksimizirali varianco vektorjev z_i.

Trditev 2.3. $\sum_{i=1}^{m} \mathrm{Var}(z_i)$ *je največja, če stolpci matrike* F *predstavljajo prvih* m *lastnih vektorjev, ki pripadajo prvim* m *največjim lastnim vrednostim kovariančne matrike* Σ*.*

Dokaz. Pišimo $\Sigma = (\Sigma_{ij})_{i,j=1}^{n,n}$, kjer je

$$\Sigma_{i,j} = E[(x_i - \mu_i)(x_j - \mu_j)].$$

Dokažimo trditev z indukcijo. Dokažimo najprej korak $m - 1 \to m$. Velja

$$\begin{aligned}
\mathrm{Var}(z_m) &= \mathrm{Var}(f_m^T X) = E[(f_m^T X)(f_m^T X)^T] - E^2[f_m^T X] \\
&= E[f_m^T X X^T f_m] - (E[f_m^T X] \cdot E[X^T f_m]) \\
&= f_m^T E[X X^T] f_m - (f_m^T E[X] \cdot E[X^T]) f_m \\
&= f_m^T (E[X X^T] - E[X] \cdot E[X^T]) f_m \\
&= f_m^T \Sigma f_m.
\end{aligned}$$

Ker je Σ kovariančna matrika, lahko naredimo lastni razcep $P\Lambda P^T$, kjer je P matrika lastnih vektorjev in Λ matrika lastnih vrednosti. Iz tega sledi:

$$f_m^T \Sigma f_m = f_m^T P \Lambda P^T f_m = w^T \Lambda w,$$

kjer je $w^T = [w_1, \ldots, w_n] = f_m^T P$. Torej velja, da je

$$\mathrm{Var}(z_m) = \sum_{i=1}^{n} w_i \lambda_i w_i = \sum_{i=1}^{n} \lambda_i w_i^2. \tag{2.18}$$

Ker velja, da je $\|f_m\| = 1$ in je P ortogonalna matrika z enotskimi stolpci, velja:

$$\begin{aligned} \|w\|^2 &= \|f_m^T P\|^2 = f_m^T P (f_m^T P)^T \\ &= f_m^T P P^T f_m = f_m^T I f_m \\ &= \|f_m\|^2 = 1. \end{aligned}$$

Po indukcijski predpostavki velja, da so bazni vektorji f_i enaki lastnim vektorjem e_i matrike Σ, za $i \in \{1, \ldots, m-1\}$. Potem sledi, da je $w_i = 0$,$i \in \{1, \ldots, m-1\}$ ker so lastni vektorji e_i pravokotni na prejšnje bazne vektorje. Torej velja:

$$\mathrm{Var}(z_m) = \sum_{i=m}^{n} \lambda_i w_i^2. \tag{2.19}$$

Očitno velja, da je $\mathrm{Var}(z_m)$ je največja, če vzamemo $w_m = 1$ in $w_j = 0$, za $j \in \{m+1 \ldots, n\}$, zaradi tega ker so lastne vrednosti razvrščene v strogo padajočem vrstnem redu $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$.
Ker je $w^T = e_m^T P$, sledi da mora biti bazni vektor f_m enak lastnemu vektorju e_m, kateremu pripada m-ta največja lastna vrednost λ_m.
Dokažimo sedaj trditev še za $m = 1$. Po istem postopku sledi:

$$\mathrm{Var}(z_1) = \sum_{i=1}^{n} \lambda_i w_i^2.$$

$\mathrm{Var}(z_1)$ je največja, če vzamemo $w_1 = 1$ in $w_j = 0$, za $j \in \{2, 3 \ldots, n\}$.
Ker je $w^T = e_1^T P$, sledi da mora biti bazni vektor f_1 enak lastnemu vektorju e_1. □

Potrebno je poudariti še, da rešitev ni enolična v dveh primerih:

- Kadar je nekaj lastnih vrednosti enakih.
- Če je ena lastna vrednost enaka 0.

Če predpostavimo, da imajo vhodne meritve X povprečje $\mu^T = (0, \ldots, 0)$, potem velja:

$$\sum_{i=1}^{n} \lambda_i = \mathrm{tr}(\Sigma) = \sum_{i=1}^{n} E[x_i^2],$$

kjer z $tr(A)$ označujemo sled matrike A. Izpeljemo pa lahko tudi sledečo enakost:

$$\mathrm{Var}(z_i) = \mathrm{Var}(f_i^T X) = \mathrm{Var}(e_i^T X) = e_i^T \Sigma e_i = e_i^T P \Lambda P^T e_i = \lambda_i. \tag{2.20}$$

Če še enkrat pogledamo enačbo (2.10) sledi, da bazo novega prostora $\mathbb{R}^m$ gradi m lastnih vektorjev matrike Σ, katerim pripada m največjih lastnih vrednosti.
Oziroma če povemo drugače, vidimo, da je rekonstrukcijska napaka r enaka vsoti $n-m$ najmanjših lastnih vrednosti neuporabljenih lastnih vektorjev:

$$\begin{aligned} r(X, \tilde{X}) &= \sum_{i=1}^{n} E[x_i^2] - \sum_{i=1}^{n} \mathrm{Var}(z_i) \\ &= \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{m} \lambda_i = \sum_{i=m+1}^{n} \lambda_i. \end{aligned}$$

Če PCA metodo uporabimo v namen faktorske analize, se poslužujemo linearnega modela faktorske analize z ortogonalnimi faktorji, kot je navedeno v članku [14]:

$$\tilde{X} = Q \cdot F, \tag{2.21}$$

kjer z F označujemo faktorje, kjer je F matrika lastnih vektorjev (baza prostora), in z U faktorske uteži, ki so koeficienti pri lastnih vrednostih za določeno meritev x_i, da velja: $x_i = q_1 f_1 + q_2 f_2 + q_3 f_3 + ... + q_m f_m$.
V magistrskem delu smo se odločili pokriti 80% varianco, kar nam je uspelo z osmimi faktorji. Varianca prostora je izračunana tako, da vsota vseh lastnih vrednosti predstavlja 100% varianco prostora.
Da bi ugotovili, kako dobro smo opisali značilke z izbranim številom faktorjev, izračunamo komunalnost značilke (ang. *communality*). Komunalnost značilke izračunamo tako, da seštejemo kvadrate uteži faktorjev, s katerimi opišemo izbrano značilko. Komunalnost pove, kolikšen delež variance značilke predstavimo z izbranim številom faktorjev. Omenjeni parameter je eden od pokazateljev, kako dobro manj dimenzionalni prostor opisuje značilke iz osnovnega prostora.
Da bi enostavneje pojasnili pomen faktorja, dobili enostavne uteži faktorja in da bi bile te pomensko razložljive, smo v magistrskem delu izvedli še rotacijo faktorjev. Uporabili smo varimax rotacijo [10]. Enostavnost uteži je definirana tako, da zadosti sledečim pogojem [23]:

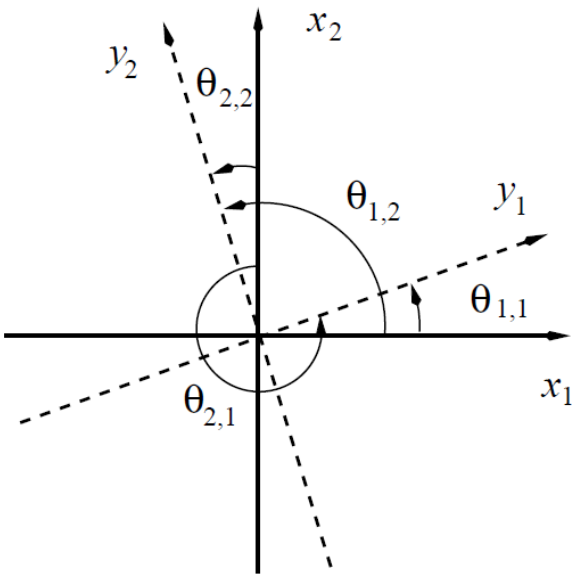
- Vsaka vrstica premore vsaj eno utež enako 0.
- Za vsak par faktorjev, obstaja velik delež uteži enak 0.
- Za vsak par faktorjev, obstaja le majhno število velikih uteži.

Da bi zadostili zgornjim pogojem, je potrebno matriko uteži matrike Q zarotirati. V magistrskem delu smo uporabili ortogonalno rotacijo $V \in R^{m\times m}$, kjer vrstice opisujejo faktorje, stolpci pa zarotirane faktorje [2]. Element $v_{i,j} \in V$ je enak kosinusu kota θ med začetnim faktorjem in zarotiranim faktorjem:

$$v_{i,j} = \cos\theta_{i,j} \tag{2.22}$$

Denimo, da imamo dva faktorja in želimo zarotirati koordinatni sistem. Zapišimo ortogonalno rotacijsko matriko V, kot kaže slika 2, za primer, ko imamo dva faktorja:

$$V = \begin{bmatrix} \cos\theta_{1,1} & \cos\theta_{1,2} \\ \cos\theta_{2,1} & \cos\theta_{2,2} \end{bmatrix} = \begin{bmatrix} \cos\theta_{1,1} & -\sin\theta_{1,1} \\ \sin\theta_{1,1} & \cos\theta_{1,1} \end{bmatrix}.$$



Slika 2: Ortogonalna rotacija V v dvorazsežnem prostoru. Kot rotacije med staro osjo j in novo zarotirano osjo i označimo kot $\theta_{i,j}$ (vzeto iz [2]).

Iz tega je jasno sklepati, da je $V^TV = I$. Rotacija V zarotira faktorske osi. Rotacijo vpeljemo v model tako, da med uteži in faktorje vrinemo matriko V in njen inverz:

$$\tilde{X} = QVV^{-1}F = Q'F',$$

pri čemer produkt QV predstavlja nove uteži Q' in produkt matrik $V^{-1}F$ predstavlja faktorje F' na zarotiranem koordinatnem sistemu.

Cilj varimax rotacije V je najti tako rešitev, da bo vsak od faktorjev imel majhno število velikih uteži in veliko število ničelnih oziroma majhnih uteži. Tako smo lažje dobili občutek, iz katerih značilk je faktor sestavljen, ter mu pripisali pomen. Varimax metoda išče tako rotacijo oziroma linearno kombinacijo pravokotnih faktorjev, da bo varianca zarotiranih uteži Q' maksimirana:

$$\nu = \left(\frac{1}{n} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} qj, i^4 - \frac{1}{n} (\sum_{j=1}^{n} q_{j,i}^2)^2 \right) \right), \tag{2.23}$$

kjer je $q_{j,i} \in Q'$ utež j-te značilke i-tega faktorja [10].
V magistrskem delu smo se posluževali opisanega postopka faktorske analize s PCA metodo in varimax rotacijo. Tako smo vsak faktor f_i prestavili z linearno kombinacijo spremenljivk x_j, katerih absolutna vrednost uteži q_i je strogo večja od 0.5. Tak način predstavitve faktorjev smo lahko uporabili zaradi uporabe varimax rotacije. Kot posledico rotacije smo dobili zelo malo velikih uteži in veliko majhnih uteži, ki smo jih zanemarili. Kakovost faktorske analize smo utemeljevali z izračunom komunalnosti spremenljivk in opisanim deležem variance manj dimenzionalnega prostora.

2.3.2 Modeliranje podatkov

Faktorje, pridobljene s faktorsko analizo, smo uporabili za modeliranje naših podatkov. Zgradili smo dva modela:

- model logistične regresije,
- regresijsko drevo.

Vsakega od modelov smo gradili dvakrat. Enkrat smo za učenje modelov uporabili vse značilke iz podatkovne zbirke, drugič pa izvedene značilke z uporabo faktorske analize.

Osnovni principi regresijske analize

Statistično modeliranje predstavi vzorec iz podatkovne zbirke z matematičnim modelom. Osnovno vodilo pravi, da se morajo vzorci čim bolj ujemati z izbranim modelom in modeli morajo biti čim bolj preprosti. V nadaljevanju smo se osredotočili na regresijsko analizo. Osnovno vodilo regresijske analize je opisati razmerje med odzivno in opisno spremenljivko.

Regresijsko analizo uporabljamo za napovedovanje, razvrščanje in pojasnjevanje zveze med podatki. Poznamo dva tipa regresijske analize: parametrično in neparametrično regresijsko analizo.
Parametrična regresijska analiza je definirana z regresijsko funkcijo. Tako opišemo razmerje med odzivno spremenljivko y in opisno spremenljivko x kot:

$$Y = f(X, \beta), \tag{2.24}$$

kjer so parametri β končni in neznani in jih je potrebno oceniti. Parametre β določimo tako, da minimiziramo napako modela:

$$\min_{\beta} \sum_{i=1}^{\ell} (y_i - f(x_i, \beta)) \,. \tag{2.25}$$

Parametrične metode regresijske analize obsegajo linearno regresijo ter vse izpeljave posplošene linearne regresije. V magistrskem delu smo kot prvi model uporabili model logistične regresije.

Logistična regresija

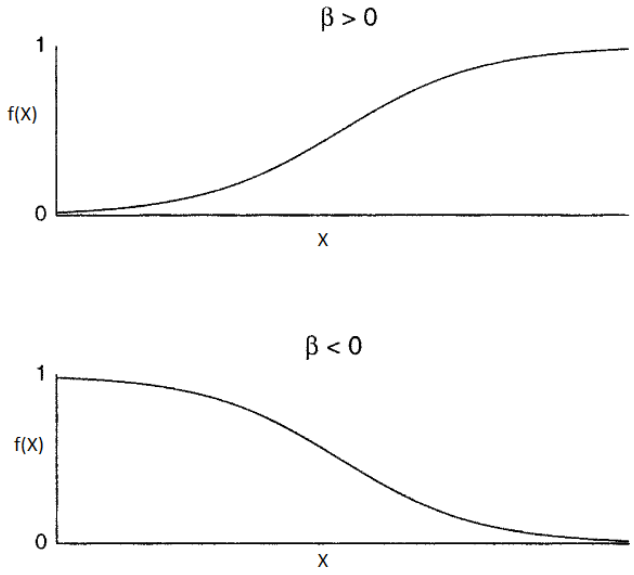
Logistična regresija je tip regresijske analize, s katero opisujemo razmerje med binarno odzivno spremenljivko in opisnimi spremenljivkami. Nelinearna razmerja med $f(X)$ in X so običajno monotona, tako da $f(X)$ monotono narašča z večanjem spremenljivke X oziroma pada. Takšen pojav lahko opišemo s krivuljo v obliki črke S (slika 3), ki jo predstavimo z zvezo:

$$Y = f(X) = \frac{e^{\beta X}}{1 + e^{\beta X}}. \tag{2.26}$$

Izpeljava za vezno funkcijo modela logistične regresije:

$$\begin{aligned} Y &= \frac{e^{\beta X}}{1 + e^{\beta X}} \\ \frac{Y}{1 - Y} &= e^{\beta X} \\ \ln \frac{Y}{1 - Y} &= \beta X. \end{aligned}$$

V našem primeru je odzivna spremenljivka odhod zavarovanca, ki ima vrednost 0 ali 1 in je binomsko porazdeljena. X predstavlja spremenljivke in faktorske spremenljivke. Odzivna spremenljivka z oznako Odhod je kategorijska z dvema razredoma. Model logistične regresije uporabimo zaradi narave odzivne spremenljivke. Izbrani model



Slika 3: Logistična funkcija glede na vrednost prametra β (vzeto iz [1]).

napoveduje binarne izide, saj so izhodne vrednosti modela na intervalu $(0, 1)$ in imajo binomsko porazdelitev. Bližje kot je izhodna vrednost modela številu ena, z večjo verjetnostjo lahko trdimo, da vzorec pripada razredu označenem z 1.

Regresijsko drevo

Drugi model, ki smo ga uporabili, je regresijsko drevo [11]. Regresijsko drevo predstavlja neparametričen model regresijske analize. Cilj regresijskega drevesa je z delitvijo podatkov na particije čim bolj natančno napovedati vrednost odzivne spremenljivke. Metoda uporablja tehniko rojenja in sicer rekurzivno deli podatkovno zbirko tako, da se bosta particiji najbolje prilagajali odzivni spremenljivki. Več smo opisali pri predstavitvi algoritma. Končni model lahko predstavimo kot dvojiško drevo.
Poznamo dva tipa odločitvenih dreves. To sta regresijsko drevo in klasifikacijsko drevo. Imata enak način delovanja, vendar sta prilagojena glede na vrsto napovedovanja. Regresijsko drevo napoveduje zvezne vrednosti odzivne spremenljivke. Klasifikacijsko drevo napoveduje diskretne vrednosti kategorijske odzivne spremenljivke. Potrebno je opozoriti, da so opisne spremenljivke lahko diskretne in zvezne. Kateri tip drevesa bomo uporabili, je odvisno izključno od odzivne spremenljivke. V magistrskem delu smo uporabili regresijsko drevo in napovedovali vrednosti odzivne spremenljivke na intervalu $[0, 1]$, kjer smo z 0 označevali aktivne zavarovance, z 1 pa zavarovance, ki niso podaljšali zavarovalne police.
V nadaljevanju smo opisali CART algoritem [4]. CART drevo je binarno odločitveno

drevo. Ideja regresijskega drevesa je določiti vejitve opisnih spremenljivk tako, da so particije podatkovne zbirke, povzročene z vejitvami, čim bolje napovedale odzivno spremenljivko.
CART algoritem [4] sledi sledečim korakom:

1. Začnemo v glavi drevesa s celotnimi podatki in izvedemo vse binarne vejitve za vsako od opisnih spremenljivk.

2. Za vsako od vejitev izračunamo povprečno vsoto kvadratov napake MSR,

$$MSR = \frac{1}{2} \sum_{c \text{ sin vozlišča } T} \left(\frac{1}{\ell} \sum_{i \in c} y_i - m_c \right)^2, \tag{2.27}$$

   kjer s T označujemo vozlišče, ki ga želimo vejit, ℓ število meritev in kjer

$$m_c = \frac{1}{\ell_c} \sum_{i \in c} y_i, \tag{2.28}$$

   predstavlja povprečno napovedano vrednost sina vejitvenega vozlišča T.

   Izberemo tisto vejitev, pri kateri je vsota kvadratov napake najmanjša. Tako vejitev imenujemo optimalna vejitev.

3. Izvedemo vse binarne vejitve za vsakega od sinov opisnih spremenljivk v notranjem vozlišču.

4. Ponavljamo drugi in tretji korak rekurzivno, dokler ne pridemo do zaustavitvenega pogoja.

Osnovno pravilo zaustavitvenega pogoja je, da prenehamo graditi drevo, kadar nadaljnje vejitve ne prinesejo izboljšave rezultata. Običajno se to zgodi, kadar delimo čista vozlišča in kadar bi z nadaljnjo vejitvijo zajeli manj kot 5% celotnih meritev podatkovne zbirke, oziroma ustavimo se, kadar z vejitvijo dosežemo prednastavljen minimalni delež meritev podatkovne zbirke.
Težava, s katero smo se srečali pri gradnji regresijskega drevesa, je težava prenasičenosti s podatki (ang. *overfitting*). Prenasičenost s podatki lahko, zaradi prevelikega prilagajanja učni množici, povzroči izjemno veliko napako, ko model razvršča nove podatke. Možni vzroki so: prevelik šum v učni množici, model je prezapleten in učna množica ni reprezentativen vzorec populacije.
Da bi se izognili prenasičenosti, se regresijsko drevo lahko obreže z določitvijo minimalnega števila meritev v listih drevesa. Večkratno grajenje drevesa nad različnimi podatki pa oceni kakovost zgrajenega modela.

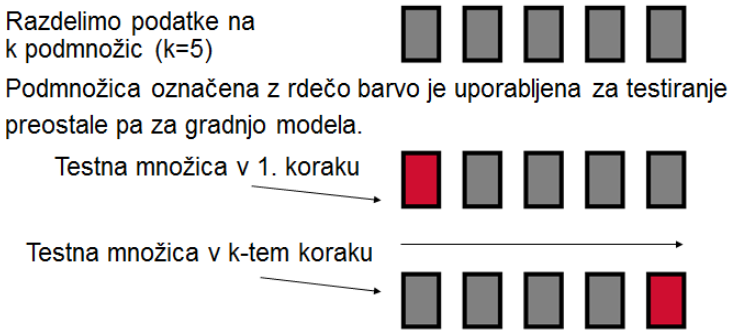
Obrezovanje drevesa poleg tega, da pripomore k posplošitvi modela, naredi model bolj preprost in kakovosten. Poznamo dva načina obrezovanja: naknadno obrezovanje (ang. *postpruning*) in vnaprejšnje rezanje (ang. *prepruning*) [24].
Vnaprejšnje rezanje je metoda sprotnega odločanja. Metoda ne reže vej, ampak za vsako novo vejitev ugotavlja, koliko bo pripomogla h kakovosti modela. Metoda je zelo uporabna, saj preprečuje nepotrebno računanje vejitev, ki bi jih z metodo naknadnega obrezovanja odstranili iz drevesa. Vendar ima metoda naknadnega obrezovanja kljub vsemu nekaj prednosti. Velikokrat se zgodi, da posamično testiranje dveh spremenljivk ne pripomore veliko h kakovosti modela. Vendar, če bi ti dve spremenljivki uporabili skupaj, bi dobili veliko boljšo kakovost modela. Metoda naknadnega obrezovanja ne dopušča, da bi se kaj takega zgodilo, saj algoritmu za grajenje drevesa pusti prosto pot, nato pa zgrajeno drevo postopoma reže od dna drevesa proti glavi drevesa.
Metoda naknadnega obrezovanja se poslužuje dveh operacij: zamenjava poddrevesa in povzdigovanje poddrevesa. Metoda zamenjava poddrevesa začne pregledovati drevo od listov proti glavi. Z vsakim pomikom proti glavi zamenjamo vejitev za list drevesa, ki nosi informacijo, kakšno odločitev smo sprejeli. Medtem ko metoda povzdigovanja drevesa vejitev na višjem nivoju drevesa zamenja tako, da povzdigne vejitev iz nižjega nivoja na njeno mesto in na novo razvrsti meritve v liste drevesa.
V magistrskem delu smo se osredotočili na regresijsko drevo. Odzivna spremenljivka vsebuje dve možni vrednosti 0 in 1, vendar bo model napovedoval vrednosti na intervalu $[0, 1]$. Zaustavitveni pogoj grajenja drevesa je bil, da je v listu drevesa vsaj 5% meritev. Pravilo vejitve je bilo določeno z minimizacijo MSR, kot je bilo opisano s CART algoritmom. Nato smo zgrajeno drevo obrezali na optimalno globino drevesa. Obrezali smo ga z metodo naknadnega obrezovanja in minimizirali izgubo navzkrižnega preverjanja.

2.4 Vrednotenje modeliranja

Da bi ocenili kakovost modela, smo uporabili metodo navzkrižnega preverjanja [8] ter napako modela predstavili z ROC krivuljo [7]. Moč napovedovanja zgrajenih modelov, ki je predstavljena z ROC krivuljami, smo testirali s statistiko analize variance, kjer smo parne primerjave naredili s post-hoc analizo po LSD statistiki (ang. *least significant difference*) [18].
Navzkrižno preverjanje je statistična metoda, namenjena razvoju in primerjavi učnih algoritmov, tako da razdelimo podatke na dva dela:

- Del, ki je namenjen učenju oziroma grajenju modela in določa parametre modela.
- Testni del, namenjen validaciji oziroma preverjanju ustreznosti modela.

Metodo navzkrižnega preverjanja uporabimo, kadar je premalo zbranih podatkov. Standardna metoda za evaluacijo se imenuje k-kratno navzkrižno preverjanje (ang. *k-fold cross-validation*). V tej metodi podatke razdelimo na k enakomerno velikih delov, katerih presek je prazna množica. Nato $k-1$ podmnožic uporabimo za učenje, 1 pa za testiranje zgrajenega modela. V vsakem koraku zamenjamo testno množico.
Slika 4 prikazuje izvajanje 5-kratnega navzkrižnega preverjanja. Množico podatkov smo razdelili na 5 enakovrednih podmnožic. Štiri podmnožice smo uporabili za grajenje modela, eno pa za testiranje. Ob vsakem koraku zamenjamo testno množico. Torej je vseh korakov natanko 5. V praksi je 10-kratno navzkrižno preverjanje dalo najboljše rezultate, zato smo tega tudi uporabili v magistrskem delu.



Slika 4: Potek izvajanja 5-kratnega navzkrižnega preverjanja.

Metoda, ki prepreči možnost pojavitve prekrivanja, se imenuje k-kratno navzkrižno preverjanje (ang. *k-fold cross-validation*). Metodo izvedemo v dveh korakih:

- celotno množico razdelimo v k podmnožic enakih velikosti,
- Nad $k-1$ podmnožic izvedemo učenje, nad preostalo pa izvedemo testiranje. Postopek ponovimo k-krat, tako da vsakič izberemo še neizbrano testno množico, kot je prikazano na sliki 4.

Končno stopnjo napake pridobimo s povprečenjem vseh iteracij.
Pri delitvi na k podmnožic je potrebno paziti na to, da bo vsaka od podmnožic reprezentativna predstavitev celotnih podatkov.
Kakovost modela pri binarni klasifikaciji se lahko oceni z ROC krivuljo (ang. *Receiver operating characteristic*). Krivulja ROC je dvodimenzionalni graf. Y os predstavlja senzitivnost, X os pa $1-$specifičnost. Senzitivnost nam pove, kolikšen delež pozitivnih primerkov smo pravilno razvrstili, medtem ko nam $1-$specifičnost pove, kolikšen delež negativnih primerkov označujemo napačno kot pozitivne. V kolikor graf vsebuje točko

$(1, 0)$, smo dosegli najboljšo binarno klasifikacijo, kar pomeni, da smo vse pozitivne primere pravilno razvrstiti in nobenega negativnega nismo napačno razvrstili. V splošnem je potrebno težiti k temu, da je ploščina pod krivuljo ROC, ki jo imenujemo AUC (ang. *area under curve*), čim bližje 1 in strogo večja od 0.5. V kolikor je ploščina pod krivuljo enaka ali manjša od 0.5, je model potrebno opustiti, saj bi enako natančnost dobili, če bi naključno določali pripadnost k razredom brez modela.
Primerjavo med modeli smo naredili tako, da smo vsakega od modelov predstavili z desetimi AUC ploščinami, ki smo jih dobili z 10-kratnim navzkrižnim preverjenem. Da bi ugotovili, ali obstajajo statistično signifikantne razlike med zgrajenimi modeli, smo uporabili testno statistiko ANOVA [18] s post-hoc analizo. Ker se je izkazalo, da se modeli med seboj statistično signifikantno razlikujejo, smo za primerjavo modelov uporabili post-hoc analizo po LSD statistiki.

3 REZULTATI

V tem razdelku smo vsako od opisanih metod uporabili na podatkovni zbirki zavarovancev in predstavili dobljene rezultate.

3.1 Analiza podatkovne zbirke

V prvem koraku analize podatkovne zbirke smo ocenili značilke, ki smo jih vključili v podatkovno zbirko. Vsako od značilk smo razdelili v dve podmnožici. V prvi so zajete vse meritve zavarovancev, ki so podaljšali avtomobilsko zavarovanje, v drugi pa so tisti zavarovanci, ki avtomobilskega zavarovanja pri opazovani zavarovalnici niso podaljšali. Cilj tega koraka je ugotoviti, ali se ti dve podmnožici statistično signifikantno razlikujeta.
Statistika, ki smo jo uporabili pri kategorijskih značilkah, je χ^2 test. Testirali smo ali obstaja signifikantna razlika med pričakovano in izmerjeno porazdelitvijo meritev med kategorijami. Zvezne značilke smo testirali z Lilliefors-ovim testom, da bi ugotovili ali sta vzorca normalno porazdeljena. Izkaže se, da porazdelitev nobena značilke ne pripada družini normalnih porazdelitev, zato v drugem koraku uporabimo Wilcoxon-ov test vsote rangov.
Ker je značilka spol kategorijska, smo za testno statistiko izbrali χ^2 test. P-vrednost testne statistike je manjša od 0.01, zato ničto hipotezo zavrnemo.
Naslednja značilka, ki smo jo testirali, je starost. Zanima nas, ali obstaja signifikantna razlika v starosti med množico oseb, ki imajo aktivno avtomobilsko zavarovanje in tistimi, ki ga nimajo. Na podlagi razultatov statistike Wilcoxon-ovega testa, ker je p-vrednost manjša od 0.01, ničto hipotezo zavrnemo.
Naslednja značilka, imenovana status, je kategorijska. Ker je p-vrednost testne statistike χ^2 enaka 0.02, ničto hipotezo zavrnemo.
Pri značilki, ki opisuje, koliko let ima zavarovanec sklenjeno avtomobilsko zavarovanje, se prav tako izkaže, da je signifikantna razlika med osebami, ki so zavarovanje prekinili, in osebami, ki imajo zavarovanje še vedno aktivno. Testna statistika Lilliefors testa zavrne ničto hipotezo, ki pravi, da je vzorec normalno porazdeljen s p-vrednostjo, manjšo od 0.01. Ker vzorca nista normalno porazdeljena, uporabimo Wilcoxonov test. Ničta hipoteza je tudi s tem testom zavrnjena in glede na to, da je p-vrednost Wilcoxon-

Tabela 2: Pregled podatkovne zbirke zavarovancev.

Značilke	Aktivni zavarovanci ($n = 22487$)	Neaktivni zavarovanci ($n = 22480$)	p-vrednost
Moški spol (%)	14754 (65.6)	14300 (63.6)	< 0.01
Starost ($\mu \pm \sigma$)	51.5 ± 13.9	48.9 ± 14.4	< 0.01
Status (%)			0.02
Brezposeln	4 (0.02)	9 (0.05)	
Zaposlen	22376 (99.5)	22395 (99.6)	
Dijak	17 (0.08)	9 (0.05)	
Študent	12 (0.05)	17 (0.08)	
Upokojenec	78 (0.35)	50 (0.22)	
Št. let avto	13.4 ± 5.7	9.2 ± 5.1	< 0.01
Št. let	16.1 ± 5.3	13.3 ± 5.9	< 0.01
Št. poti	2.7 ± 1.3	2.6 ± 1.3	< 0.01
Št. stebrov	2.6 ± 1.0	2.3 ± 1.1	< 0.01
Št. škod	0.22 ± 0.2	0.24 ± 0.4	< 0.01
Št. polic	2.2 ± 1.9	1.7 ± 1.2	< 0.01
Št. avto. polic	1.1 ± 0.2	1.1 ± 0.8	< 0.01
Premija	693 ± 117	519 ± 121	< 0.01
Avto. premija	370 ± 165	325 ± 314	< 0.01
Škode	402 ± 117	360 ± 123	< 0.01
Avto. škode	181 ± 742	270 ± 156	< 0.01
Naklon	0.71 ± 1.6	0.51 ± 1.43	< 0.01
Leta škod	0.14 ± 0.15	0.16 ± 0.24	< 0.01

ovega testa manjša od 0.01 lahko sklepamo, da se množici statistično signifikantno razlikujeta.

Da bi ugotovili, ali je dejstvo, da stranka sklepa police preko raznih prodajnih kanalov, vplivalo na odhod stranke, smo tudi to lastnost vpeljali v model. Zanima pa nas, ali obstaja signifikantna razlika v količini prodajnih kanalov med osebami, ki so ostale aktivne in osebami, ki v tem trenutku nimajo aktivnega avtomobilskega zavarovanja. Wilcoxonov test zavrne ničto hipotezo s p-vrednostjo, manjšo od 0.01 in potrdi, da obstaja signifikantna razlika med medianama rangov vzorcev.

Naslednja značilka na podoben način poskuša ugotoviti, ali je zvestost in razpršenost zavarovanj ključnega pomena, da zavarovanec ohrani aktivno polico. Značilka prešteje,

v koliko zavarovalnih stebrih je zavarovanec imel sklenjeno zavarovanje. P-vrednost Wilcoxon-ovega testa je bila manjša od 0.01, iz česar sledi da lahko zavrnemo ničto hipotezo.

Pri značilki, ki prešteje leta, v katerih je zavarovanec imel aktivno vsaj eno zavarovanje, pričakujemo podobne rezultate kot pri značilki, ki je štela aktivna leta v avtomobilskem stebru zavarovanj. In testna statistika Wilcoxon-ovega testa znova zavrne ničto hipotezo, ki pravi, da sta vzorca aktivnih in neaktivnih oseb enako porazdeljena, s p-vrednostjo manjšo od 0.01.

Naslednjih 5 značilk bo poskušalo ugotoviti možnost nezadovoljstva med zavarovanci. Predpostavka, iz katere lahko razberemo nezadovoljstvo je, da so bili zavarovanci, ki so bolj pogosto imeli škode, nezadovoljni z uslugami zavarovalnice in so zaradi tega neaktivni. Vse značilke somo testirali z Lilliefors testom in s p-vrednostjo, manjšo od 0.01, se zavrne ničta hipoteza, da je katerikoli izmed vzorcev značilk normalno porazdeljen.

Prva med petimi značilkami je značilka, ki šteje, koliko škod na eni polici je zavarovanec v povprečju na leto uveljavljal. S statistiko Wilcoxon-ovega testa zavrnemo ničto hipotezo, da sta mediani vzorcev enaki.

Druga izmed petih značilk je likvidirana škoda, ki smo jo v povprečju izplačali zavarovancu na polico. Rezultat Wilcoxon-ovega testa nam da vedeti, da je očitna razlika med osebami, ki so aktivne, in tistimi, ki so prekinile zavarovanje z p-vrednostjo manjšo od 0.01.

Tretja od petih značilk, ki obravnavajo škodne dogodke, je delež škodnih let. Zanima nas, ali obstaja signifikantna razlika med aktivnimi in neaktivnimi osebami v deležu škodnih let. Preverimo, kakšno ugotovitev smo dobili s testno statistiko. P-vrednost Wilcoxon-ovega testa je manjša od 0.01. Sprejmemo alternativno hipotezo, da se mediana prvega vzorca statistično signifikantno razlikuje od mediane drugega vzorca.

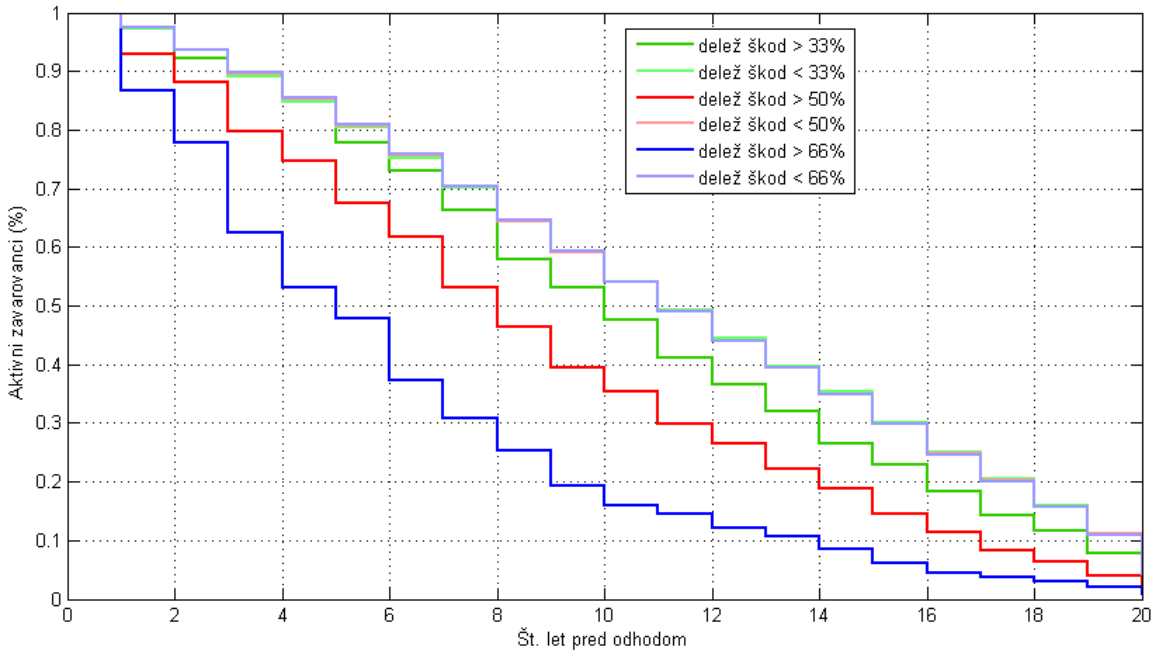
Značilka, ki bi lahko veliko povedala o naravi in lastnostih zavarovancev, je, koliko likvidiranih škod v povprečju zavarovancu zavarovalnica izplača na leto. Wilcoxon-ov test je zavrnil ničto hipotezo, da imata vzorca enako porazdelitev s p-vrednostjo, manjšo od 0.01.

Zadnja značilka, ki opisuje škodno obnašanje zavarovanca, je dinamična značilka, imenovana naklon. Značilka opisuje trend upadanja oziroma naraščanja uveljavljanja škod. Zanima nas, ali je med zavarovanci, ki niso podaljšali zavarovanja, veliko tistih, katerim je trend uveljavljanja škod proti koncu narasel. Wilcoxon-ov test je nakazal na signifikantno razliko vzorcev, kjer je bila p-vrednost manjša od 0.01.

Naslednji značilki, ki smo ju analizirali, opisujeta vsoto vplačanih premij v avtomobilsko oziroma vsa zavarovanja na zavarovalnici. Wilcoxon-ov test je v obeh primerih zavrnil ničto hipotezo z p-vrednostjo manjšo od 0.01, da sta si vzorca enaka.

Zadnja izmed šestnajstih opisnih značilk je povprečje sklenjenih polic na leto. Značilka nam bo poskušala odgovoriti na vprašanje, ali se količina polic pri osebah, ki so avtomobilsko zavarovanje podaljšali, razlikuje od količine polic, ki niso podaljšali avtomobilskega zavarovanja. Uporabili smo enostranski Wilcoxon-ov test. Alternativna hipoteza v primeru sprejetja potrdi razliko med medianami in sicer mediana vzorca oseb, ki so še vedno aktivne, je večja kot mediana oseb, ki niso več aktivne. Testna statistika zavrne ničto hipotez pri p-vrednosti manjši od 0.01.

3.2 Časovna analiza odhodov zavarovancev



Slika 5: Krivulje preživetja za zavarovance, razdeljene po kriteriju deleža škodnih let.

Z analizo preživetja, ki uporabi Kaplan-Meierjevo neparametrično metodo [13], smo predstavili čas zavarovancev do odhoda. Zavarovance, ki so prekinili zavarovanje, smo razdelili v dve skupini. Parameter, s katerim smo zavarovance razvrščali, je bil delež škodnih let. Analizo smo ponovili trikrat. V prvem koraku smo v eno skupino razvrstili vse osebe, ki so vsako tretje leto ali pogosteje imele kakšen škodni dogodek. V drugo skupino pa osebe, ki so škodne dogodke upoštevale manj. Opaziti je, da delež oseb, ki so bile bolj izpostavljene škodnim, hitreje upada v primerjavi z drugo skupino, ki škodnih let praktično ni imela. Primerjali smo še osebe, ki so imele kakšen škodni dogodek vsaj vsako drugo leto z osebami, ki so imela manj škodnih let. In kot tretjo primerjavo smo razdelili podatkovno zbirko na zavarovance, ki so bile le vsako tretje leto

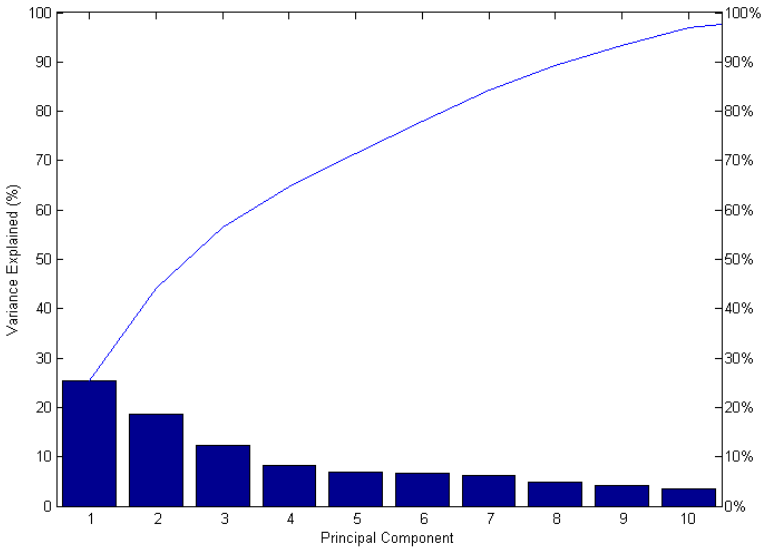
brez škodnega dogodka in na tiste, ki so bile manj izpostavljene škodnim dogodkom. Slika 5 prikaže krivulje preživetja za vsako od pripadajočih skupin. Opaziti je večji osip zavarovancev v skupinah, kjer smo zajeli zavarovance, ki so izpostavljeni škodnim dogodkom. Trend hitrejšega upadanja je viden že pri zavarovancih, ki so imeli vsako tretje leto škodni dogodek, v primerjavi z njimi nasprotno skupino, ki je imela delež škodnih dogodkov manjši od 33%. Z večanjem deleža škodnih let je jasno opaziti, da se vsako leto delež zavarovancev občutno zmanjša. Krivulja preživetja za zavarovance, ki so so imeli delež škodnih dogodkov večji od 66%, doseže 50% osip zavarovancev po petih letih, medtem ko krivulja komplementarne skupine potrebuje 11 let, da doseže 50% osip zavarovancev. Osip zavarovancev lahko pripišemo nezadovoljstvu, kumuliranju škodnih točk in posledično manjšanju bonusov, kar je povzročilo višje premije.
Skupine zavarovancev, ki opisujejo zavarovance z manjšim deležem škodnih let, opisuje linearna krivulja, pri čemer ne moremo izpostaviti nobene kritične točke. V povprečju je vsako leto osip zavarovancev 5%, kar je lahko posledica tržne konkurence oziroma naključnih dogodkov.

3.3 Faktorska analiza

V magistrskem delu smo uporabili faktorsko analizo s PCA metodo in varimax rotacijo, da bi ugotovili povezanost med opisnimi značilkami. V faktorsko analizo smo vključili vseh štirinajst numeričnih značilk in dobili štirinajst faktorjev. V nadaljnjo analizo smo vzeli prvih osem faktorjev, saj ti pokrijejo malo več kot 80% variance prostora. Kumulativna varianca faktorjev in varianca prvih desetih faktorjev je predstavljena s tako imenovanim Paretto grafom (slika 6). Poznamo več smernic, koliko faktorjev izbrati. Faktorje uredimo od najvišjega do najnižjega, glede na opisano varianco prostora.

- Izberemo prvih n, ki skupaj opišejo vsaj 80% variance.
- Zadnji zajeti faktor mora opisati vsaj 5% variance.
- Število faktorjev določimo glede na grafični prikaz. Izrišemo razmerje med opisano varianco in številom faktorjev. "Koleno" grafa nam nakaže na ustrezno število faktorjev.

V magistrskem delu smo se odločili za določitev števila faktorjev, tako da zajamemo vsaj 80% variance prvotnega prostora.
Kakovost modela lahko preverimo tudi s preverjanjem komunalnosti značilk. Tabela 3 prikaže, kolikšen delež značilke opišemo z izbranim številom faktorjev.
Kot je opaziti, je model najslabše opisal značilke, ki predstavljajo količino sklenjenih polic ter leta pri zavarovalnici in avtomobilskem zavarovanju.



Slika 6: Graf opisanega deleža variance prostora.

Tabela 3: Tabela opisuje, kolikšen delež značilke opišemo z osmimi faktorji.

Naklon	98.4%
Škoda za avto. zav	97.2%
Vplačane premije	96.4%
Likvidirane škode	96.2%
Št. prodajnih kanalov	92.4%
Št. škod	91.5%
Delež škodnih let	90.4%
Starost:	90.3%
Št. tipov zavarovanj	89.3%
Premija za avto. zav.	85.3%
Leta na zavarovalnici	82.7%
Leta pri avto. zav.	78.5%
Št. avto polic	75.6%
Št. polic	70.9%

Prvi faktor, ki opiše največji delež variance, lahko opišemo kot stalnost zavarovanca. Značilki, ki sestavljata prvi faktor, sta število stebrov zavarovanj, v katere je vključen zavarovanec in število let pri zavarovalnici. Enostavno se lahko razloži, da več let kot smo na zavarovalnici, več raznolikih zavarovanj sklenemo. Drugi faktor predstavi po-

vezavo med značilkama škodnih dogodkov. Več škodnih dogodkov kot imamo, večji je delež škodnih let. Faktor bi lahko poimenovali kot škodno tveganje. Tretji faktor opiše povezavo med vplačanimi premijami in škodami. Faktor opisuje izpostavljenost škodnim izplačilom. Večje kot je vplačilo premije na polico, večje je tudi tveganje za škodni dogodek. Četrti faktor je opisan z eno dominantno značilko, ki predstavlja likvidirane škode avtomobilskih zavarovanj. Če podrobneje pregledamo faktor, ugotovimo, da je nanj, sicer z zelo majhno utežjo, vezano število zavarovančevih let na avtomobilskem zavarovanju, kar jasno opiše izpostavljenost škodnih dogodkov. Peti faktor opisuje povezavo med vplačanimi premijami in številom sklenjenih polic, kar nakaže na premoženjski status. Šesti faktor dominantno opisuje značilka naklon. Faktor predstavlja trend rasti uveljavljanja škod. Sedmi faktor opisuje povezavo med menjavo prodajnih kanalov glede na leta zavarovanja. Lahko ga opišemo kot fleksibilnost stranke pri načinu sklepanja zavarovanj. Osmi faktor opisuje pričakovano dobo podaljševanja zavarovanj glede na starostno skupino zavarovanca.
Struktura faktorjev z uteženimi značilkami je natančneje predstavljena v tabeli 4.

Tabela 4: Tabela opisuje, kako je značilka utežena za posamezen faktor.

Faktor1 :	0.925 stTipZavarovanj + 0.618 letaAS
Faktor2 :	0.924 stSkod + 0.918 letaSkod
Faktor3 :	0.947 premije + 0.912 skode
Faktor4 :	0.935 skodaAvto
Faktor5 :	(-0.845) premijaAvto + (-0.618) stPolic
Faktor6 :	0.966 naklon
Faktor7 :	0.534 letaAvto + 0.956 stProdajnihKanalov
Faktor8 :	(-0.949) starost + (-0.546) letaAvto + (-0.524) letaAS

3.4 Analiza modela

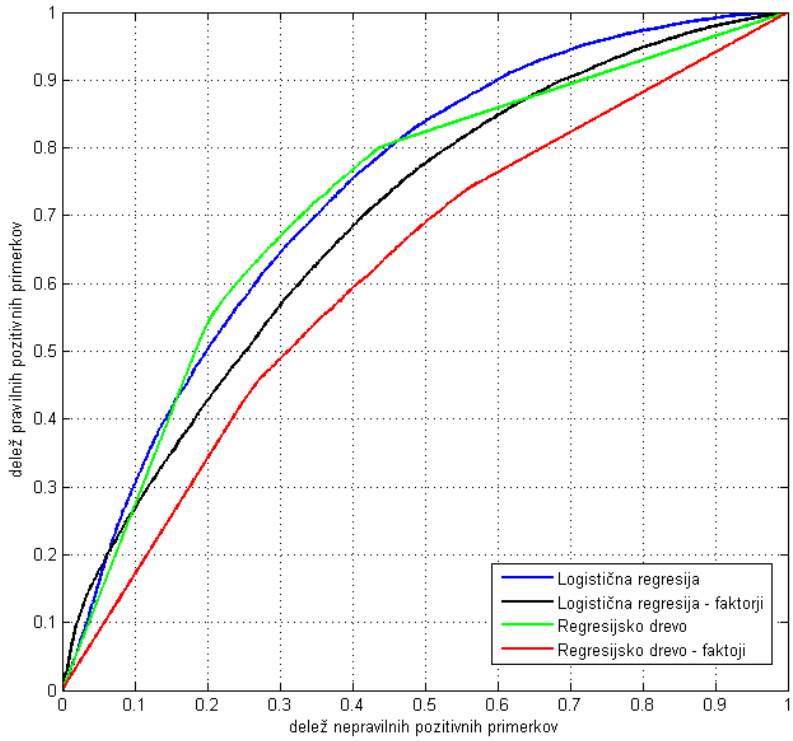
V magistrskem delu smo za napovedovanje odhoda zavarovancev uporabili dva modela: model logistične regresije in regresijsko drevo.
Vsakega od modelov smo gradili dvakrat. Prvič smo v model vključili vse značilke, saj smo z analizo podatkovne zbirke ugotovili, da je pri vsaki značilki signifikantna razlika med osebami, ki so ostale zavarovane in osebami, ki niso podaljšale zavarovanja. Podatkovno zbirko smo razdelili na 10 enako velikih podmnožic. Ena podmnožica je bila uporabljena za testiranje modela, ostalih devet pa smo uporabili za učenje modela. Da bi ugotovili, ali obstajajo razlike med množicami, smo gradnjo modela ponovili

desetkrat ter v vsakem koraku izbrali drugo testno množico. Tako smo dobili 10 ROC krivulj za en model. Na podlagi dobljenih krivulj smo izrisali povprečno ROC krivuljo. Ko smo model gradili drugič, smo celoten postopek ponovili, le zamenjali smo tip podatkov, ki smo jih vključili v model. V drugem koraku smo namesto značilk podatkovne zbirke modelirali faktorje, pridobljene s faktorsko analizo.

Tabela 5: Tabela kakovosti zgrajenih modelov.

Model	Povprečna AUC ploščina $\pm\ \sigma$
Logistična regresija	0.738 ± 0.006
Logistična regresija s faktorji	0.694 ± 0.009
Regresijsko drevo	0.724 ± 0.008
Regresijsko drevo s faktorji	0.621 ± 0.004

Poleg ROC krivulje nam veliko o kvaliteti modela pove tudi AUC ploščina pod krivuljo. Zgrajeni model ima povprečno ploščino AUC enako 0.738. Poglejmo še, kako je model zgrajen in katere značilke največ pripomorejo pri gradnji modela. Značilke razvrstimo po p-vrednostih od najmanjše do največje. Značilke, ki imajo p-vrednost manjšo od 0.05, imajo velik vpliv na model. Med prvimi po p-vrednosti sodijo: število let pri avtomobilskem zavarovanju, število zavarovalnih stebrov, v katerih ima zavarovanec aktivno polico ter število let na zavarovalnici. Značilki, ki nista signifikantno značilni za model, sta spol in status.
V naslednjem koraku smo ponovili celoten postopek gradnje modela s faktorji.
Dobljena rezultata modelov najlažje primerjamo, če povprečni ROC krivulji izrišemo na eni sliki. Zelo enostavno se opazi razlika tudi v primerjavi AUC ploščin. AUC ploščine drugega modela so na intervalu $[0.683, 0.699]$.
Povprečna AUC ploščina drugega modela je 0.694. Poglejmo še, kakšne p-vrednosti je model dodelil faktorjem. Vsi faktorji, razen tretjega in četrtega, imajo p-vrednosti tako majhne, da jih označujemo z 0. Tretji in četrti faktor imata p-vrednost večjo od 0.3, kar pomeni, da ne vplivata na model. Iz tega lahko sledi, da vsota vplačanih premij in skupna vsota izplačanih škod ne vplivata na odhod zavarovanca.
Naslednji model, s katerim smo se srečali, je regresijsko drevo. Regresijsko drevo je neparametrični model, pri katerem ne predpostavljamo zvezo med opisnimi in odzivnimi značilkami v obliki funkcije, ampak iščemo vzorce v obliki pravil iz vhodnih podatkov. Postopek grajenja je potekal na enak način, kot pri logistični regresiji. V prvem koraku smo gradili model iz značilk, ki smo jih zajeli v podatkovni zbirki.
Povprečna AUC ploščina je 0.724. Značilke, ki so najbolj diskriminatorne in najbolje delijo množico na particije, so na vrhu drevesa. Med prvimi tremi značilkami je število



Slika 7: ROC krivulje zgrajenih modelov.

let z aktivno polico pri avtomobilskem zavarovanju, povprečna vplačana premija in značilka, ki šteje, v kolikšnih zavarovalniških stebrih ima zavarovanec aktivno polico. Sedaj smo regresijsko drevo uporabili za modeliranje faktorjev. ROC krivulje regresijskega drevesa, ki je grajen iz faktorjev, so nižje kot pri modelu, grajenim iz značilk. Povprečna AUC ploščina pa je 0.621. Faktor 1, ki je opisan z značilkama število aktivnih let na zavarovalnici in značilko, ki šteje v kolikšnih zavarovalniških stebrih ima zavarovanec aktivno polico, je glava drevesa, kar pomeni, da največ pripomore pri gradnji modela.

Če primerjamo povprečne ploščine modelov, bi lahko rekli, da obstaja razlika med njimi, vendar vprašamo se lahko, ali je razlika statistično signifikantna. Na to vprašanje smo si odgovorili v razdelku, ki sledi.

Natančneje povedano, preverili smo, ali obstajajo statistično signifikantne razlike med kakovostjo modelov.

Za testno statistiko smo uporabili enosmerno analizo variance oz. ANOVA. S testno statistiko ANOVA smo ugotovili, da je p-vrednost manjša od 0.01. Iz tega lahko na podlagi AUC ploščin sklepamo, da se modeli napovedovanja odhoda zavarovanca statistično signifikantno razlikujejo.

Da bi ugotovili, kateri modeli se razlikujejo in katere lahko posplošimo, smo uporabili postopke večkratnih primerjav (ang. *multiple comparison methods*). S post-hoc analizo s kriterijem LSD smo ugotovili, da se vse AUC vrednosti med različnimi modeli paroma statistično signifikantno razlikujejo, s čimer lahko sklepamo, da z različnimi modeli dobimo različne rezultate napovedovanja odhodov zavarovancev. Iz grafov ROC (slika 7) in tabele AUC (tabela 5) lahko ugotovimo, da najboljše rezultate dobimo z modelom logistične regresije.

4 RAZPRAVA

Rezultate, ki smo jih predstavili v predhodnem poglavju, smo obrazložili in pojasnili, kako vplivajo na trg zavarovalniških poslov. Iz dobljenih modelov smo predstavili tipe zavarovancev.

V razdelku analiza podatkovne zbirke smo z analizo preživetja ocenjevali, kolikšen bo upad zavarovancev iz avtomobilskega zavarovanja glede na število škodnih let. Ugotovili smo, da so zavarovanci, kateri vsaj vsako tretje leto upoštevajo škodni dogodek, veliko bolj podvrženi odhodu kot preostali zavarovanci. Z zavarovalniškega vidika lahko sklepamo, da je posledica odhoda pripisovanje malusov, kar povzroči referentu nezmožnost nižanja premije in nezmožnost uveljavljanja komercialnih in izrednih popustov, kar posledično pomeni višjo premijo. Zaradi tega zavarovanci raje odidejo h konkurenci, saj ta nima vpogleda v zgodovino škod in jim ponudi popuste za novo sklenjeno zavarovanje, na podlagi premijskega razreda v katerem se oseba nahaja. Predvidevamo, da so taki zavarovanci nestalni in bolj naklonjeni menjavam zavarovalnic. V letnem poročilu zavarovalnice [26] je bilo tudi navedeno, da zavarovalnica zdaj oblikuje avtomobilska zavarovanja po meri stranke, prav tako pa oblikuje premijo glede na rizičnost posameznih manjših segmentov zavarovancev, pri čemer upošteva številne dejavnike, med njimi pa tudi starost voznika in kraj stalnega bivališča, kar se s krivuljo preživetja tudi opazi.

Opaziti je tudi, da zavarovalnica za vsako zavarovančevo leto izgubi 5% zavarovancev. Sklepamo lahko, da je to posledica konkurence in ekonomske krize, saj se fakturirana premija kasko zavarovanj iz leta v leto znižuje in je bila 12% nižja od premije v letu 2012. Tudi pri avtomobilskih zavarovanjih se odraža ekonomska kriza, saj se ljudje vedno manj odločajo za zavarovanje drugega avtomobila, ki ga še imajo. Trg novih vozil je sicer ohranil raven prodaje, vendar prodajalci menijo, da je to večinoma na račun enodnevnih registracij vozil, ki se izvozijo v tujino. Nekateri prodajalci vozil govorijo tudi o 20% do 30% manjšanju prodaje vozil, kar se zelo pozna tudi na zavarovanosti. Izpad premije pri kasko zavarovanju je rezultat hude konkurence in upada števila zavarovanj [26].

V naslednjem koraku smo z modeli logistične regresije in regresijskim drevesom gradili model iz značilk in faktorjev. Modele smo uporabili za napovedovanje odhodov strank. Testna statistika LSD je pokazala, da je najbolj kakovosten model logistične regresije,

ki za vhodne podatke uporabi značilke podatkovne zbirke.

Model logistične regresije je kot najpomembnejše značilke označil starost, leta pri avtomobilskem zavarovanju, število prodajnih poti, naklon, povprečje vseh škod, število zavarovalnih stebrov, v katere je zavarovanec vključen, število let na zavarovalnici. Značilkam število let pri avtomobilskem zavarovanju in število zavarovalnih stebrov, v katere je zavarovanec vključen, je pripisal izjemno visok negativni parameter. Iz tega parametra lahko zaključimo sledeče: z vsako dodatno enoto parametra se zmanjša verjetnost, da bi zavarovanec odšel od zavarovalnice. Vzemimo za primer, da imamo dva zavarovanca, ki se razlikujeta samo v eni enoti značilke število zavarovalniških stebrov. Zaključimo lahko, da zavarovanec, ki bo imel večjo vrednost te značilke, je za 0.6-krat manj naklonjen k odhodu kot zavarovanec, ki ima nižjo vrednost značilke št. zavarovalniških stebrov. Enako lahko sklepamo za značilko število let pri avtomobilskem zavarovanju, le da se pri tej značilki z vsakim letom zmanjša možnost odhoda za 0.8-krat. Značilka, ki je nakazala na večje tveganje odhoda, je naklon. V kolikor bi imeli dva zavarovanca z enakimi vrednostmi značilk, bi oseba z večjim naklonom imela za 1.2-krat več možnosti za odhod.

Povzamemo lahko, da je model nakazal na pričakovane rezultate ter nakazal, da so osebe, ki so manj podvržene škodnim dogodkom, osebe, ki imajo dlje časa sklenjeno zavarovaje in osebe, katerih zavarovalne police so razpršene na več zavarovalnih stebrov, manj naklonjene k odhodu.

Kot manj učinkovit se je izkazal model, ki je za vhodne podatke vzel faktorje. Kot smo ugotovili z logističnim modelom, ki je za vhodne podatke vzel značilke iz podatkovne zbirke, sta se število let pri zavarovalnici in število let pri avtomobilskem zavarovanju, izkazala za zelo pomembna parametra. S pregledom komunalnosti smo opazili, da je faktorska analiza ravno ta dva parametra najslabše opisala, zato lahko predpostavimo, da je ena od posledic upada kakovosti ravno v tem. Vendar vsekakor je model logistične regresije s faktorji določil enake karakteristike, kot predhodno opisan model. Prvi faktor, kateri nakazuje na stalnost stranke, je označen kot najpomembnejši faktor. Nakaže, da ob primerjavi zavarovancev in ob predpostavki, da je vrednost preostalih faktorjev enaka, ima za 0.6-krat manjšo možnost prekinitve police zavarovanec, ki ima večje vrednosti faktorja stalnost. Medtem ko ob enaki predpostavki, pri analiziranju drugega faktorja, ugotovimo, da je večja izpostavljenost škodnim dogodkom poveča za 1.3-krat možnosti odhoda.

Če razdelimo modele glede na vhodne podatke, se je model regresijskega drevesa v obeh primerih izkazal s slabšo kakovostjo napovedovanja. Vendar zanimivo je opaziti, da sta tudi ta dva modela izpostavila enake lastnosti zavarovanca. Model, katerega vhodni podatki so bile značilke, je za najpomembnejši značilki pri klasifikaciji uporabnikov določil število let pri avtomobilskem zavarovanju in delež škodnih let. Prav tako je

model, grajen s faktorji, označil prvi faktor in peti faktor za najpomembnejša faktorja. Spomnimo se, da prvi faktor označuje stalnost stranke, peti faktor pa količino sklenjenih zavarovanj.

Če se sprehodimo skozi celotno analizo, lahko zaključimo, da so osebe, izpostavljene pogostejšim škodnim zahtevkom, nestalne in je njihova časovna premica za polovico krajša od oseb, ki imajo manjši delež škodnih let. Z gradnjo modelov smo ugotovili, da je stalnost zavarovanca opisana s številom let pri zavarovalnici in razpršenostjo zavarovalnih polic med ostale zavarovalne produkte, ključnega pomena pri ohranjanju stranke. V bodoče bi lahko predlagali tržno akcijo, ki bi nagovarjala zavarovance avtomobilskih zavarovanj k sklenitvi dodatnega zavarovanja, ki bi stranko vključila v nov zavarovalniški steber.

Pomembno je poudariti, da lahko uporabljen način modeliranja podatkov uporabimo za reševanje različnih binarnih vprašanj, ne glede na področje raziskovanja. Potrebno je razumeti delovanje statističnih metod, znati interpretirati dobljene rezultate ter jih umestiti v raziskovano področje.

5 ZAKLJUČEK

Zmožnost napovedati, s kolikšno verjetnostjo bo obstoječa stranka odšla h konkurenci oziroma ocenjevati lojalnost stranke, je neprecenljiva za razvoj podjetja in uspeh na trgu. Prednost je boljše razumevanje strank in njenih potreb. Če razumemo potrebe strank, lahko oddelek za stranke na podlagi ugotovljenih vzorcev prilagodi ponudbo. Zavedati pa se moramo, da je potrebno konstantno spremljati dogajanje, saj en poseg ne odpravi težave. Potrebe povpraševalcev po storitvah se s časom spreminjajo, potrebno je izvajanje akcij vseskozi prilagajati novim potrebam in razmeram na trgu, pri tem pa nam pomagajo različne analize.

Glede na današnje stanje trga, njegovo raznolikost in kompleksnost, je zelo težko s prostim očesom zaznati, kaj naj bi bil vzorec, zaradi katerega povzročamo nezadovoljstvo med strankami. S tem modelom lahko zaznamo ključne dejavnike, ki vplivajo na zadovoljstvo in lojalnost stranke in zmanjšamo stroške tržne akcije, saj nagovorimo le tiste stranke, na katere zaznani dejavniki najbolj vplivajo. V splošnem izvajamo tržno akcijo na neko ciljno skupino. Prednost take akcije je boljši odziv na akcijo, manjši stroški akcije, saj nagovorimo manjše število strank, in prilagajanje akcije profilu stranke tako, da se stranki dejansko ponudi točno to, kar potrebuje.

V magistrskem delu smo predstavili model, ki ga lahko uporabimo na raznih področjih in ne zgolj na področju avtomobilskih zavarovanj oziroma trženja. Za razumevanje rezultatov, ki jih model izračuna, je potrebno teoretično znanje o modelu in strokovno znanje raziskovanega področja.

Literatura

[1] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, Second Edition, 2002.

[2] H. Abdi, Factor Rotations in Factor Analyses. V *The University of Texas*, Dallas, 2003.

[3] C. Archaux, A. Martin in A. Khenchaf An SVM Based Churn Detector in Prepaid Mobile Telephony. V *International Conference on Information and Communication Technologies (ICTTA),Damas*, 2004, 19–23.

[4] L. Breiman, J.H. Friedman, R. Olshen in C.J. Stone, *Classification and Regression Tree*, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific California, 1984.

[5] I. Bose in X. Chen, Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn, *Journal of organizational Computing and Electronic Commerce* 19 (2009), 133–151.

[6] R. Chengalvarayan in L. Deng, Use of generalized dynamic feature parameters for speech recognition, *IEEE Trans. Speech and Audio Processing* 5(3) (1997), 232–242.

[7] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters* 27(8) (2006), 861–874.

[8] E. Frank in I. Witten, *Data Mining-Practical Machine Learning Tools And Techniques*, Morgan Kaufmann publishers, San Francisco, 2005.

[9] M. Guillen, J. Parner, C. Densgsoe in A.M. Perez-Martin Customer Loyalty in the Insurence Industry. V *Conference in Actuarial Science and FInance on Samos*, 2002, 20–22.

[10] H. H. Harman, *Modern Factor Analysis*, University of Chicago Press, Third Edition, 1976.

[11] T. Hastie, R. Tibshirani in J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Second Edition, 2008.

[12] A.H. Karp, Using Logistic Regression to Predict Customer Retention. V *The NorthEast SAS Users Group*, (NESUG), 1998.

[13] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, Hoboken, NJ: Wiley-Interscience, 2002.

[14] J.N. Lawley in A.E. Maxwell, *Factor Analysis as a Statistical Method*, American Elsevier Publishing Co., 1971.

[15] A. Lemmens in C. Croux, Bagging and Boosting Classification Trees to Predict Churn, *Journal of Marketing research* 43(2) (2006), 276–286.

[16] H.W. Lilliefors, On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown, *Journal of the American Statistical Association* 62 (1967), 399–402.

[17] A. A. Miranda, Y-A. Borgne in G. Bontempi, New Routes from Minimal Approximation Error to Principal Components, *Neural Processing Letters* 27(3) (2008), 197–207.

[18] D. C. Montgomery in G. C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, Third Edition, 2002.

[19] E.W.T. Ngai in L. Xiu, Application of Data Mining Techniques in Customer Relationship Management: A literature review and classification, *Expert Systems with Applications* 36 (2009), 2592–2602.

[20] B. Ratner, *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*, Taylor & Francis Group, Second Edition, 2011.

[21] J. Su, K. Cooper,T. Robinson in B. Jordan Customer Retention Predictive Modeling in HealthCare Insurance Industry. V *Proc. 17th annual SouthEast SAS Users Group*, Birmingham, 2007.

[22] S. Theodoridis in K. Koutroumbas, *Pattern Recognition*, Academic Press, Third Edition, 2006.

[23] L. L. Thurstone, *Multiple Factor Analysis*, University of Chicago Press, 1947.

[24] I. H. Witten in E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann publications, Second Edition, 2005.

[25] Y. Xie, L. Xiu in E.W.T. Ngai, Customer Churn Prediction Using Improved Balanced Random Forest, *Expert Systems with Applications* 36 (2009), 5445–5449.

[26] Elektronsko letno poročilo AS 2013
URL: http://www.as-skupina.si/financno-sredisce/letna-porocila
(26.2.2015)